PRO GRADU -TUTKIELMA

Hanna Lindholm

Tilasto-ohjelma Pivotti

Pivotti toimii MS Excel-ohjelmassa ja hyödyntää Excelin PivotTablea

TAMPEREEN YLIOPISTO
Informaatiotieteiden yksikkö
Tilastotiede

Marraskuu 2013

Tampereen yliopisto

Informaatiotieteiden yksikkö

LINDHOLM, HANNA: Tilasto-ohjelma Pivotti
Pivotti toimii MS Excel-ohjelmassa ja hyödyntää Excelin PivotTablea

Pro gradu -tutkielma, 25 s., 10 liites.

Tilastotiede

Marraskuu 2013

Tiivistelmä

Tilasto-ohjelma Pivotti sisältää käyttöliittymän Excelin PivotTableen ja se hyödyntää PivotTablea taulukoiden ja kuvioiden tekemisessä. PivotTable ei ole kokemattomille käyttäjille tuttu eikä helppokäyttöinen, joten Pivotti on heille avuksi.

Pivotilla pystyy piirtämään ympyrä- ja pylväsdiagrammin, histogrammin sekä pisteparven. Histogrammiin on tehty automaattinen luokitus, mutta käyttäjä voi halutessaan tehdä luokittelun myös manuaalisesti. Yksiulotteisesta jakaumasta voidaan laskea prosentit ja tunnuslukuja. Pivotilla voi laskea seuraavat tunnusluvut: keskiarvo, mediaani, moodi, minimi, maksimi, keskihajonta, varianssi, vaihteluväli, ala- ja yläkvartiili sekä prosenttipisteet. Ohjelmalla voi myös tehdä ristiintaulukon, piirtää siitä kuvan ja laskea χ^2-riippumattomuustestin sekä teoreettiset frekvenssit. Kaikkiin tarkasteluihin on mahdollista lisätä ehdollistava muuttuja. Pivotti on laajennettavissa käsittämään muitakin tilastollisia menetelmiä.

Tutkielma toimii manuaalina Pivotin käyttöön ja siinä kerrotaan yksityiskohtaisesti, mitä tilastomenetelmiä ohjelmasta löytyy ja mistä valikoista ne löytyvät. Ohjelman käyttöä havainnollistetaan myös kuvien avulla. Tutkielman avulla kokematonkin käyttäjä pystyy käyttämään Pivottia sekä kertaamaan käytettävien tilastomenetelmien teoriaa. Lisäksi tutkielmassa kerrotaan ohjelman sisältämistä tarkistuksista.

Lopuksi Pivotin toimintaa ja osa-alueita havainnollistetaan esimerkkiaineiston avulla. Tämän esimerkkiaineiston avulla näytetään käytännön tilanteessa, millaisia kuvaajia ja taulukoita ohjelmalla pystyy tulostamaan, sekä mitä tunnuslukuja sillä saa laskettua. Lisäksi ohjelman antamat tulostukset tulkitaan.

Sisältö

1 Johdanto

1.1 Mikä Pivotti on?

Microsoft Officen Excelissä on olemassa valmiina hyödyllinen taulukkotyökalu, jota sanotaan PivotTableksi. PivotTablella voi muun muassa tehdä yksiulotteisia tai moniulotteisia jakaumia, käyttää yhtä tai useampaa ehdollistavaa muuttujaa ja esittää arvot lukumäärinä, prosentteina ja summana. Kuitenkin PivotTablen luominen ja käyttäminen on kokemattomalle Excelin käyttäjälle todennäköisesti hankalaa, koska sen käyttö ei ole erityisen helppoa. Koska Excelin käyttö on hyvin yleistä ja PivotTable on hyödyllinen työkalu, olen tehnyt sen käyttöä helpottamaan tilasto-ohjelman nimeltä Pivotti.

Pivotti on käyttöliittymä PivotTableen ja helpottaa siten huomattavasti tämän käyttämistä. Lisäksi Pivotti sisältää joitakin tilastollisia perusanalysointimenetelmiä aineiston tarkastelua varten. Ohjelma on tehty suomen kielellä, ja se on mahdollisimman helppokäyttöinen ja huomioi näin kokemattomatkin käyttäjät. Pivotti on toteutettu Visual Basic -ohjelmointikielellä (VBA) MS Excelin versiolla 2010, mutta se on yhteensopiva myös aiempien versioiden kanssa. Pivotti on muokattu sanasta PivotTable. Nimi on suomalaiselle helppo lausua ja taivuttaa, ja se kuvaa ohjelman pääsisältöä.

1.2 Pivotin toiminnot ja tavoite

Kun Pivotti käynnistetään, se tekee PivotTable-taulukon. Sen jälkeen käyttäjä saa valita lomakkeelta, mitä muuta tehdään. Ensimmäisenä vaihtoehtona on yksiulotteinen jakauma. Siitä voi lisäksi laskea prosentit, kumulatiiviset prosentit ja erilaisia tunnuslukuja. Toisena vaihtoehtona on grafiikkavalikko, joka sisältää ympyrädiagrammin, pylväsdiagrammin, histogrammin ja pisteparven. Excelissä histogrammin kuvaaja ei ole yhtä helposti löydettävissä kuin muu grafiikka, joten kokemattomalle käyttäjälle Pivotti on suureksi avuksi histogrammin tekemisessä. Kolmas vaihtoehto on ristiintaulukko, josta voidaan lisäksi laskea prosentit riveittäin, sarakkeittain tai yhteensä sekä χ^2-riippumattomuustesti. Lisäksi ristiintaulukosta voidaan vielä piirtää vaakapalkkikuvaaja. Kaikkien näiden taulukoiden ja kuvaajien tekemisessä hyödynnetään alussa tehtyä PivotTablea. Lisäksi PivotTableen voi lisätä ehdollistavan muuttujan, jolloin kaikki toiminnot tehdään sen määrittämissä luokissa.

Kuten edellä kerrotusta käy ilmi, Pivottiin on toteutettu vain muutamia tilastollisia menetelmiä. Ohjelmaa on kuitenkin mahdollista laajentaa ja liittää siihen esimerkiksi regressioanalyysi, varianssianalyysi tai useampiulotteisia ristiintaulukoita. Ohjelman varsinainen tavoite — auttaa kokemattomia käyttäjiä käyttämään PivotTablea ja huomaamaan siinä olevat mahdollisuudet — täyttyy kuitenkin jo näillä peruselementeillä.

1.3 Pivotti suhteessa muihin tilasto-ohjelmiin

Jos Pivottia verrataan muihin tilasto-ohjelmiin, se muistuttaa eniten Tixeliä, joka on tilastotieteen professori Pentti Mannisen tekemä ohjelma. Myös Tixel toimii Excel-ympäristössä, se on helppokäyttöinen ja saatavissa suomen kielellä. Tixelissä on kuitenkin enemmän toimintoja kuin Pivotissa. Pivotin erityispiirteenä puolestaan on pivotTablen käyttö.

Joihinkin Pivotin valikoihin on otettu mallia SPSS-ohjelmasta. SPSS-ohjelmassa kaikki grafiikka löytyy yhdestä valikosta ja samaan on pyritty myös Pivotissa. Poikkeuksena tästä on ristiintaulukon kuvaaja, joka löytyy vain ristiintaulukon yhteydestä. Tähän ratkaisuun on päädytty, jotta ristiintaulukon oikeanlainen kuvaaja löytyisi mahdollisimman helposti. Esimerkiksi SPSS-ohjelmassa vastaavaa kuvaaja ei löydy yhtä helposti. Toisaalta tähän on syynä sekin, että SPSS:ssä on paljon enemmän vaihtoehtoja erilaisille kuvaajille kuin Pivotissa. Toinen kohta, jossa mallia on otettu SPSS-ohjelmasta, on Pivotin tunnuslukuvalikko. SPSS:ssä monet valikot toimivat periaatteella, jossa valittava muuttuja siirretään toiseen listaan tai ruutuun. Tämän tapaista menetelmää on käytetty Pivotissa prosenttipisteiden valinnan yhteydessä. Ensimmäiseen ruutuun kirjoitetaan haluttu prosenttipiste, joka sitten siirretään toiseen ruutuun, johon kerätään lista kaikista halutuista prosenttipisteistä.

Suuri osa Pivotin toiminnoista hyödyntää jo olemassa olevia Excelin funktioita tai muita toimintoja. Pivotissa ne ovat kuitenkin helposti löydettävissä ja samalla kertaa voi laskea esimerkiksi useita erilaisia tunnuslukuja sen sijaan, että ne kaikki etsittäisiin erikseen. Tämä säästää aikaa ja vaivaa.

1.4 Pivotin ohjelmoinnista

Pivotti sisältää karkeasti arvioiden noin 140 sivua VBA-koodia. Koodi on kirjoitettu hyviä ohjelmointitapoja noudattaen, se on sisennetty ja kommentoitu. Pivottia on tarkistettu huolellisesti, jotta se ei jumiutuisi missään tilanteessa, mutta on silti mahdollista, että jokin erikoinen tilanne on jäänyt huomiotta. Kuten arvata saattaa, ohjelmointi on ollut työlästä. Ohjelman tekoon kulunutta tarkkaa aikaa on mahdoton sanoa, mutta sadoista tunneista on kysymys. Kerrotaan vielä lyhyesti, miksi ohjelmointi on näin työlästä.

Ohjelmointi etenee siten, että ensin tekijä miettii, mitä hän haluaa tehdä ja kirjoittaa vastaavan koodin. Tämä vaihe voi olla nopeakin. Seuraavaksi testataan tekeekö koodi sen, mitä oli suunniteltu. Sen jälkeen huomataan mahdolliset virheet ja puutteet, korjataan koodia ja testataan uudelleen. Tätä korjaamisen ja testaamisen vuorottelua jatketaan, kunnes ohjelma toimii kuten on suunniteltu. Varsinkin tarkistaminen ja virheiden etsiminen ovat aikaa vieviä osuuksia, koska virheiden löytäminen ei ole yksinkertaista. Lisäksi hyvin pienetkin asiat voivat olla yllättävän monimutkaisia. Pelkästään oikean rivin laskeminen tulostukselle on joskus yllättävän työlästä, varsinkin jos tarkastelussa on mukana ehdollistava muuttuja. Havainnollistetaan asiaa esimerkillä. Halutaan tulostaa teksti Khi^ 2 jokaiselle ehdollistavan muuttujan arvolle, koska kaikille näille arvoille lasketaan χ^2-testisuure viereiseen soluun.

Lisäksi tulostetaan muitakin tietoja ristiintaulukosta ja mahdollisesti teoreettiset frekvenssit, joten Khi^ 2 tekstien väliin täytyy jättää useita rivejä. Tällöin halutun rivin indeksi saadaan kaavalla:

$(15 + rivLkm) * k + 3$, missä

$rivLkm$ = PivotTable-taulukon rivien lukumäärä
k = monesko ehdollistavan muuttujan luokka on kyseessä

Edellinen esimerkki oli vielä suhteellisen yksinkertainen. Monimutkaisemmaksi tilanne muuttuu, kun tulostetaan ristiintaulukon absoluuttiset frekvenssit kaikissa ehdollistavan muuttujan luokissa. Ongelmana on varsinkin tilanne, jossa jokin ehdollistavan muuttujan luokista ei saa kaikkia sarake- tai rivimuuttujan arvoja. Koodissa pitää siis varautua kaikenlaisiin tilanteisiin, jotta arvot tulostuisivat oikeille riveille ja sarakkeille. Tämä luonnollisesti vaatii paljon testaamista, aikaa ja pohdiskelua.

Tässä työssä ei käsitellä enempää Pivotin ohjelmointia. Jos joku haluaa lisätietoja, niitä voi kysyä tekijältä. Pivotin koodi ei ole tämän työn liitteenä, mutta myös se on saatavissa tekijältä.

1.5 Tutkielman rakenne

Tämä tutkielma toimii manuaalina Pivotin käyttöä varten. Tavoitteena on, että käyttäjä saisi tutkielman avulla käsityksen menetelmien teoriasta, käyttöohjeet ohjelmaan, yleiskäsityksen siitä, mitä kaikkea ohjelmalla voidaan tehdä ja miten saatujen tulosten tulkinta etenee. Jos siis käyttäjällä on perustiedot tilastotieteestä, hän pystyy tämän manuaalin avulla kertaamaan teoriaa, käyttämään ohjelmaa ja tulkitsemaan tuloksia. Ensin kerrotaan ohjelman sisältämien tilastollisten menetelmien teoriaa (luku 2). Sen jälkeen esitellään ohjelman käyttöliittymä ja valikot sekä analyysimenetelmät, grafiikka ja muut toiminnot (luku 3).

Neljännessä luvussa esitellään ohjelman toimintaa vielä esimerkkiaineiston avulla. Luku etenee samalla tavoin kuin tilastolliset tutkielmat. Ensin esitellään aineisto grafiikan ja tunnuslukujen avulla ja sen jälkeen edetään riippuvuustarkasteluihin. Esimerkkiaineistona on käytetty yleisesti opetuskäytössä ollutta, ympäristömielipiteitä koskevaa aineistoa (Ympäristö 2000). Osa aineiston kyselylomakkeesta on liitteenä. Koko kyselylomake on sivulla http://www.fsd.uta.fi/fi/aineistot/luettelo/FSD0115/quF0115_fin.pdf. Tutkielman viimeinen luku (luku 5) on yhteenveto.

2 Ohjelmassa käytettyjen tilastollisten menetelmien teoriaa

2.1 Yksiulotteinen jakauma

Yksiulotteinen jakauma eli frekvenssijakauma muodostetaan jakamalla muuttujan kaikki mahdolliset arvot luokkiin $E_1, E_2, \ldots, E_k$ ja ilmoittamalla kuinka monta havaintoyksikköä kuuluu kuhunkin luokkaan. Tiettyyn luokkaan kuuluvien havaintojen lukumäärää kutsutaan frekvenssiksi. Merkintä f_i tarkoittaa i. luokan havaintojen lukumäärää.

Tavallisen frekvenssin rinnalla määritellään usein myös seuraavat frekvenssit:
Prosenttiosuus: $f_i\% = 100 \cdot (f_i/n)$
Summafrekvenssi: $F_i = f_1 + f_2 + \cdots + f_i$
Kumulatiiviset prosentit:
$Fi\% = 100 \cdot (f_1 + f_2 + \cdots + f_i)/n$

Kaikissa kaavoissa i viittaa tarkasteltavaan luokkaan E_i ja n havaintojen kokonaislukumäärään (otoksessa tai populaatiossa).

Tunnusluvut

Tunnusluku on arvo, jonka avulla pyritään kuvaamaan jakauman ominaisuuksia. Tunnuslukuja voidaan laskea sekä populaatiolle, että otokselle. Tässä esitetyt tunnusluvut on tarkoitettu otokselle, mutta niiden avulla voidaan estimoida populaation vastaavia tunnuslukuja. Käytettävän tunnusluvun valintaa rajoittaa käytetty mittaustaso. Myös jakauman muoto tulisi huomioida, kun tulkitaan tunnuslukuja.

Tyyppiarvo eli **moodi** on se arvo, joka esiintyy useimmiten (eli jolla on suurin frekvenssi). Se ei kuitenkaan aina ole yksikäsitteinen; usealla luokalla saattaa olla sama suurin frekvenssi. Moodi ei välttämättä ole jakauman keskellä ja jatkuvalla muuttujalla se ei ole kovin käyttökelpoinen.

Mediaani on keskimmäinen arvo, kun muuttujan arvot on järjestetty suuruusjärjestykseen. Jos muuttujalla on parillinen määrä arvoja, mediaani on kahden keskimmäisen arvon keskiarvo. Muuttujan arvoista 50 % on suurempia kuin mediaani ja 50 % muuttujan arvoista on pienempiä kuin mediaani. Mediaani ei ole herkkä poikkeaville arvoille ja on usein erittäin käyttökelpoinen vinojen jakaumien tapauksessa.

Keskiarvo on erittäin herkkä poikkeaville arvoille varsinkin, jos aineisto on suhteellisen pieni. Tällöin mediaani saattaa olla parempi vaihtoehto keskiluvuksi. Keskiarvo lasketaan seuraavasti: Muuttujan x arvot ovat havaintoyksiköittäin muotoa $x_1, x_2, \ldots, x_n$. Tällöin keskiarvo $\bar{x}$

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Alakvartiili on luku, joka jakaa muuttujan arvot kahtia siten, että 25 % arvoista on pienempiä kuin alakvartiili. **Yläkvartiili** on vastaavasti luku, joka jakaa muuttujan arvot kahtia siten, että 75 % arvoista on pienempiä kuin yläkvartiili ts. 25 % arvoista on suurempia kuin yläkvartiili. Alakvartiili, mediaani ja yläkvartiili siis jakavat muuttujan arvot neljään havaintomääriltään yhtä suureen osaan.

Prosenttipisteitä laskettaessa valitaan haluttu prosentti p. Nyt p:n prosenttipiste on arvo, jota pienempiä on p % muuttujan arvoista. Esimerkiksi alakvartiili on 25 %:n prosenttipiste. Kiinnostuksen kohteena ovat yleensä jakauman ääripäät.

Vaihteluväli saadaan, kun vähennetään muuttujan suurimmasta arvosta (**maksimista**) muuttujan pienin arvo (**minimi**).

Keskihajonta mittaa sitä, kuinka tiiviisti muuttujan havainnot ovat keskittyneet sen keskiarvon ympärille. Ääritapauksessa, eli silloin kun kaikki havainnot ovat yhtä suuria, keskihajonta on nolla. Muulloin keskihajonta on aina suurempi kuin nolla. Keskihajonta s on (otos)varianssin s^2 neliöjuuri. **Varianssi** määritellään seuraavasti:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2$$

Keskihajonta on tällöin $s = \sqrt{s^2}$. Laskettaessa populaation varianssia jaetaan summa n:llä, $(n-1)$:n asemesta.

2.2 Kuvaajatyypit

Ympyrädiagrammi eli sektoridiagrammi eli piirakkakuvio on ympyrän muotoinen kuvio, jossa eri väreillä tai eri kuvioinneilla tehdyt sektorit edustavat muuttujan eri arvoja. Mitä suurempi sektori, sitä enemmän kyseistä arvoa esiintyy. Ympyrädiagrammi korostaa hyvin eri arvojen osuuksia suhteessa toisiinsa. Ympyrädiagrammia ei kannata käyttää, jos muuttujan arvoja on monia.

Pylväsdiagrammi piirtää kuvan, jossa pylvään pituus kertoo arvon suhteellisen osuuden prosentteina (tai arvon frekvenssin). Pylväsdiagrammi sopii käytettäväksi tapauksissa, joissa muuttuja ei ole jatkuva tai tasavälinen luokitus ei ole mahdollinen. Pylväät voivat olla pysty- tai vaakatasossa. Vaakapalkkeja käytetään yleensä epätasavälisen luokituksen tapauksessa. Ympyrä- ja pylväsdiagrammit sopivat hyvin kvalitatiivisille muuttujille.

Histogrammi piirretään siten, että se muodostuu suorakulmioista, joiden leveys on luokan pituus, korkeus (tasavälisessä luokituksessa) luokan frekvenssi ja kantojen kärkipisteet ovat todellisissa luokkarajoissa. Histogrammi voidaan muodostaa, kun muuttuja on jatkuva. Frekvenssihistogrammi kuvaa jakauman muotoa.

Kaikki edellä mainitut kuvaajatyypit sopivat hyvin yksiulotteisen jakauman tapauksiin. Kahden kvantitatiivisen muuttujan riippuvuuden tutkimiseen sopii **pisteparvi** eli korrelaatiodiagrammi eli hajontakuvio. Se muodostetaan piirtämällä koor-

dinaatistoon pisteet $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ siten, että selitettävä muuttuja on y-akselilla ja selittävä muuttuja x-akselilla. Pisteparvi antaa yleensä hyvän yleiskuvan muuttujien mahdollisesta riippuvuudesta.

2.3 Ristiintaulukko ja χ^2-riippumattomuustesti

2.3.1 Ristiintaulukko

Kahden kvalitatiivisen muuttujan tapauksessa riippuvuutta tarkastellaan vertailemalla ehdollisia prosenttijakaumia keskenään. Jos jakaumat poikkeavat toisistaan, sanotaan selitettävän muuttujan y riippuvan selittäjästä eli x-muuttujasta. Jos jakaumat ovat lähes samanlaiset, niin riippuvuutta ei ole. Jos tilanne halutaan otoksen perusteella yleistää koko populaatioon, tilastollinen testaus on tarpeen. Testinä käytetään χ^2-riippumattomuustestiä.

Ristiintaulukon eli kaksiulotteisen frekvenssijakauman muodostaminen: jaetaan muuttujan x arvot luokkiin $E_1, E_2, \ldots, E_k$ ja muuttujan y arvot luokkiin $F_1, F_2, \ldots, F_r$. Nyt ristiintaulukko on muotoa:

Taulukko 2.1. Ristiintaulukon muodostaminen.

	E_1	E_2	$\ldots$	E_k	
F_1	f_{11}	f_{12}	$\ldots$	f_{1k}	$f_{1\bullet}$
F_2	f_{21}	f_{22}	$\ldots$	f_{2k}	$f_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
F_r	f_{r1}	f_{r2}	$\ldots$	f_{rk}	$f_{r\bullet}$
	$f_{\bullet 1}$	$f_{\bullet 2}$	$\ldots$	$f_{\bullet k}$	$f_{\bullet\bullet}$

Ristiintaulukossa (2.1) f_{ij} tarkoittaa solun F_i, E_j havaintomäärää eli solufrekvenssiä, $f_{\bullet j}$ tarkoittaa j. sarakkeen frekvenssien summaa, $f_{i\bullet}$ puolestaan i. rivin frekvenssien summaa ja $f_{\bullet\bullet}$ kaikkien frekvenssien summaa, eli havaintomäärää n. Reunafrekvenssit $f_{1\bullet}, f_{2\bullet}, \ldots, f_{r\bullet}$ muodostavat muuttujan y frekvenssijakauman ja reunafrekvenssit $f_{\bullet 1}, f_{\bullet 2}, \ldots, f_{\bullet k}$ muodostavat muuttujan x frekvenssijakauman. Sarakkeiden frekvenssit muodostavat y:n ehdollisen jakauman. Edellä on muodostettu $r \times k$ -frekvenssitaulukko. Jos taulukko on kokoa 2×2 , sitä kutsutaan nelikentäksi. Sopiva tapa kuvata kahden kategorisen muuttujan välistä riippuvuutta graafisesti on vierekkäiset pystypylväät tai peräkkäiset vaakapalkit.

2.3.2 χ^2-riippumattomuustesti

Riippuvuuden merkitsevyys ristiintaulukossa voidaan testata χ^2-testisuureen avulla. Hypoteesit χ^2-riippumattomuustestissä ovat:

H_0: muuttujat x ja y ovat riippumattomia.
H_1: muuttujat x ja y eivät ole riippumattomia.

Testisuureen laskemista varten tarvitaan solujen havaitut frekvenssit f_{ij} sekä odotetut eli teoreettiset frekvenssit e_{ij}. Odotetut frekvenssit lasketaan havaittujen frekvenssien ja otoskoon n avulla. Kaava odotettujen frekvenssien laskemiseen on muotoa:

$$e_{ij} = \frac{f_{i\bullet} f_{\bullet j}}{n}$$

Testisuureen kaava on

$$\sum_{i=1}^{r} \sum_{j=1}^{k} \frac{(f_{ij} - e_{ij})^2}{e_{ij}}.$$

Kun H_0 on tosi, niin testisuure noudattaa asymptoottisesti χ^2-jakaumaa vapausasteilla df, missä $df = (r-1) \cdot (k-1)$.

Jotta testiä voitaisiin käyttää, täytyy seuraavien ehtojen olla voimassa:
1) Jos $df > 1$, niin kaikkien $e_{ij} > 1$ ja enintään 20 % saa olla < 5.
2) Jos $df = 1$ ja $20 \leq n \leq 40$, niin kaikkien $e_{ij} \leq 5$. Jos $df = 1$ ja $n > 40$, niin testiä voi käyttää ilman rajoitteita.

Yleisesti katsotaan, että näiden ehtojen täyttyessä jakauma noudattaa riittävän tarkasti χ^2-jakaumaa. χ^2-testisuureen arvon perusteella saadaan p-arvo, joka kertoo pienimmän riskitason, jolla H_0 voidaan hylätä. Tarvitaan kuitenkin ristiintaulukko ja siihen ehdolliset prosenttijakaumat selittäjän mukaan, ennen kuin riippuvuuden ilmenemistapa voidaan päätellä. Tästä syystä myös Pivottiin on lisätty mahdollisuudet laskea ehdolliset prosenttijakaumat.

3 Pivotin esittely

Pivotin käyttämisen oletuksena on, että käsiteltävä data on havaintomatriisimuodossa. Lisäksi datan pitää alkaa työarkin ensimmäisestä sarakkeesta ja riviltä kolme siten, että riville kolme on kirjoitettu muuttujien nimet. Datan katsotaan loppuvan, kun vastaan tulee täysin tyhjä rivi tai sarake. Muuttujien arvot voidaan kirjoittaa numeroina tai tekstinä. Jos muuttujalle kirjoitetaan selitteitä, niin ne kirjoitetaan kyseisen muuttujan sarakkeelle datan alapuolelle siten, että datan ja selitteiden väliin jää yksi tyhjä rivi. Jos selitteitä on kirjoitettu, ohjelma käsittelee muuttujaa vastedes tekstitietoa sisältävänä muuttujana. Pivotti ei piirrä histogrammia tai pisteparvea tekstitietoa sisältäville muuttujille, mutta sen sijaan se ilmoittaa, että muuttujassa on tekstitietoa. Jos muuttujaa kuitenkin haluttaisiin käsitellä numeerisena ja piirtää esimerkiksi pisteparvi, voidaan selitteet poistaa ja suorittaa haluttu toiminto sen jälkeen. Datassa olevia tyhjiä soluja käsitellään puuttuvana tietona. Tunnusluvut ja prosentit lasketaan ilman puuttuvaa tietoa. Poikkeuksena on tilanne, jossa ehdollistavassa muuttujassa on puuttuvaa tietoa. Tällöin lasketaan prosentit ja tunnusluvut ilman ehtomuuttujan puuttuvaa tietoa ja sitten koko muuttujasta.

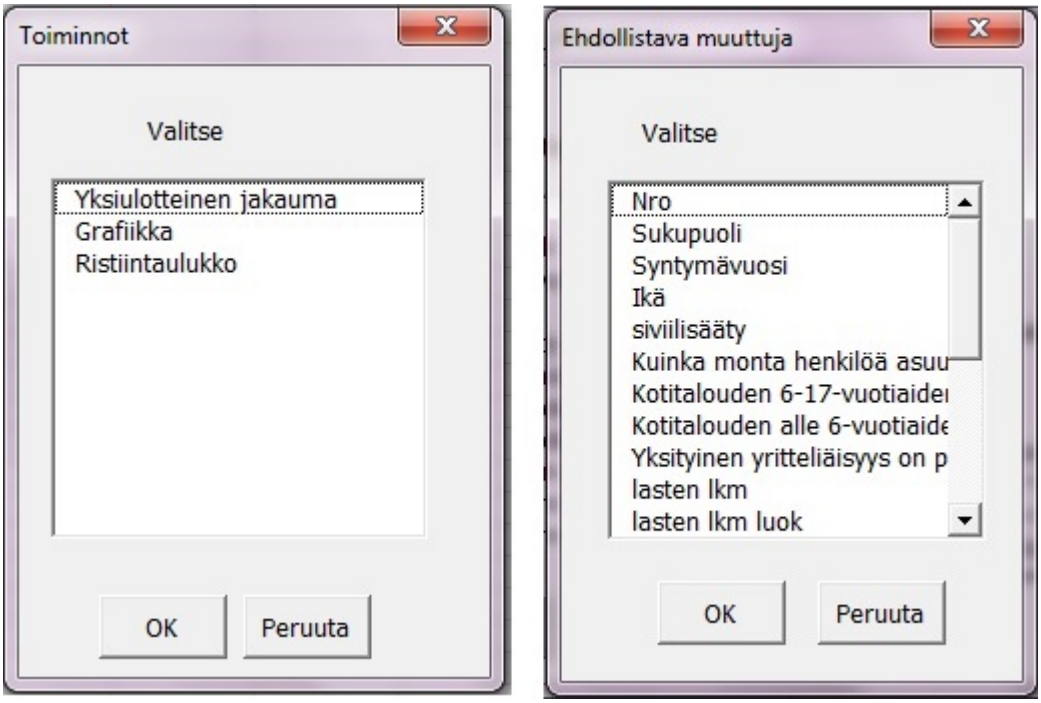
3.1 Päävalikko

Kun ohjelma käynnistetään, tulee näkyviin päävalikko. Valikossa on vaihtoehtoina yksiulotteinen jakauma, grafiikka ja ristiintaulukko. Toiminnoista voi valita vain yhden kerrallaan. Jos käyttäjä ei ole valinnut näistä mitään, ohjelma kehottaa valitsemaan toiminnon. Jos painetaan Peruuta-painiketta, valikko sulkeutuu. Kaikissa Pivotin lomakkeissa on OK-painike ja Peruuta-painike. Ensiksi mainitusta siirrytään eteenpäin ja Peruuta-toiminnolla lomake sulkeutuu. Kun tehdään yksiulotteinen jakauma, grafiikkaa tai ristiintaulukko, ohjelma avaa uuden työkirjan, jonka työarkeille PivotTable-taulukko, tunnusluvut ja kuvaajat tulostetaan. Kun Pivotti käynnistetään ja päävalikko tulee näkyviin, ohjelma muodostaa PivotTable-taulukon uudelle työarkille. Tätä PivotTablea hyödynnetään myöhemmin, kun muodostetaan jakaumia.

Kaikkiin toimintoihin on lisätty tarkistuksia, jotta ohjelman suoritus ei päättyisi kesken ja jotta käyttäjä saisi palautetta vääristä valinnoista. Kaikissa toiminnoissa, joihin liittyy muuttujan valintaa, ohjelma ilmoittaa, jos mitään muuttujaa ei ole valittu, eikä se siirry eteenpäin ennen kuin muuttuja on valittu. Poikkeuksena tästä on ehdollistavan muuttujan valinta, sillä tätä muuttujaa ei aina haluta valita. Jos ristiintaulukkoon tai pisteparveen on valittu kaksi samaa muuttujaa, ohjelma huomauttaa ja siirtyy eteenpäin vasta, kun käyttäjä valitsee kaksi eri muuttujaa. Jos käyttäjä valitsee ehdollistavaan muuttujaan saman muuttujan kuin varsinaiseen toimintoon, niin ohjelma huomauttaa, etteivät valittava muuttuja ja ehdollistava muuttuja voi olla samoja, ja se suorittaa toiminnon ilman ehdollistavaa muuttujaa. Ohjelmassa on estetty tyhjän muuttujan käyttäminen. Tyhjällä muuttujalla tarkoitetaan muuttujaa, jolla ei ole muita tietoja kuin nimi. Jos käyttäjä yrittää valita tällaisen muuttujan, ohjelma

huomauttaa asiasta eikä etene ennen kuin käyttäjä on valinnut jonkin toisen muuttujan. Kuviossa 3.1 on esitetty päävalikko ja ehdollistavan muuttujan valikko. Jos ehtomuuttujan arvoja on enemmän kuin sata, ohjelma huomauttaa tästä eikä anna valita sitä ehtomuuttujaksi. Näin toimitaan siksi, että vanhempien Excelin versioiden sarakkeet eivät riitä, jos ehtomuuttuja saa liikaa arvoja.

Pivotti on tehty vain yhdelle ehdollistavalle muuttujalle, mutta se on laajennettavissa usealle ehdollistavalle muuttujalle. Toinen vaihtoehto on, että käyttäjä muodostaa uuden muuttujan, jossa on useampi ehtomuuttuja yhdessä. Joka tapauksessa on siis mahdollista tarkastella myös useita ehtomuuttujia samalla kertaa.

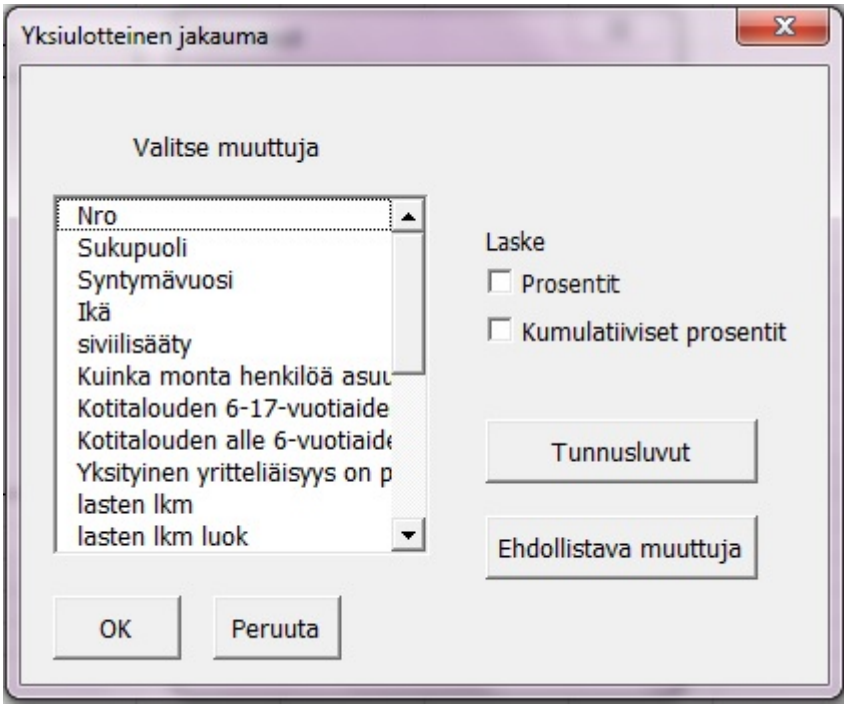


Kuvio 3.1. Pivotin päävalikko ja ehdollistavan muuttujan valikko.

3.2 Yksiulotteinen jakauma

Yksiulotteisen jakauman valikosta (kuvio 3.2) voi valita muuttujan, jonka frekvenssijakauma esitetään PivotTable-taulukossa. Lisäksi voidaan valita muuttujan prosenttijakauma, kumulatiiviset prosentit, ehdollistava muuttuja ja laskea erilaisia tunnuslukuja. Tunnusluvuista voidaan valita keskiarvo, mediaani, moodi, minimi, maksimi, keskihajonta, varianssi, vaihteluväli, ala- ja yläkvartiili sekä käyttäjän määrittelemät prosenttipisteet. Prosenttipisteet kirjoitetaan vasemmalla olevaan pieneen ruutuun desimaalilukuna siten, että ne sijaitsevat välillä [0,1]. Kun prosenttipiste on kirjoitettu, käyttäjä painaa painiketta, jossa on merkkijono >>>>. Tällöin arvo siirtyy oikealla puolella olevaan listausvalikkoon. Jos halutaan lisää prosenttipisteitä, kirjoitetaan taas uusi arvo ja se siirretään listaan. Käyttäjä voi valita niin monta prosenttipistettä kuin hän haluaa. Jos jokin prosenttipiste halutaan pois listasta, valitaan

se ja klikataan painiketta, jossa on merkkijono <<<<. Sen jälkeen prosenttipiste poistuu listasta ja sitä ei lasketa. Jos käyttäjä syöttää kirjaimia tai luvun, joka ei ole välillä [0,1], niin ohjelma kehottaa häntä muuttamaan syötettä. Kuviossa 3.3 on esitetty valittavien tunnuslukujen valikko.



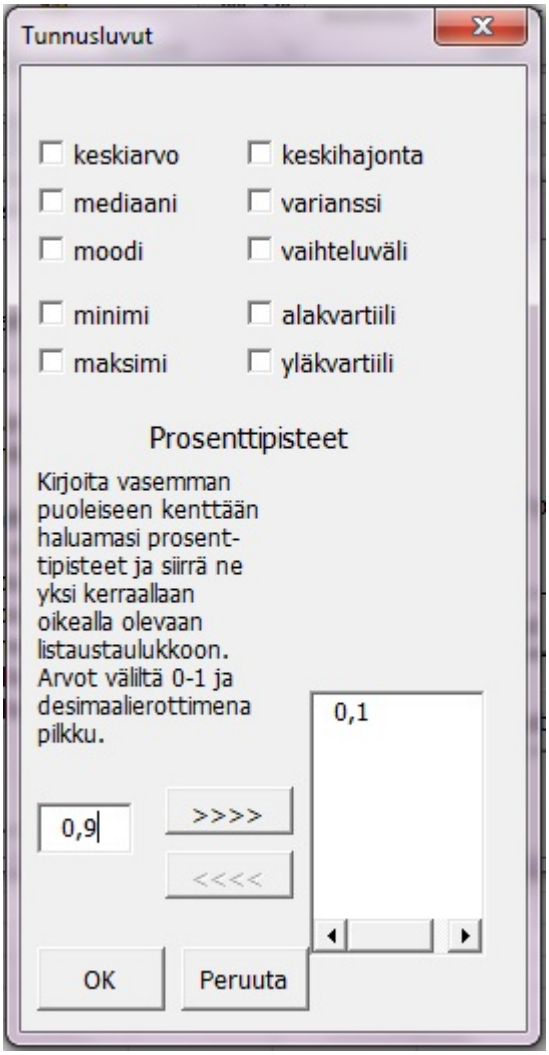
Kuvio 3.2. Yksiulotteisen jakauman valikko.

Ohjelma tekee ensin muuttujan frekvenssijakauman PivotTableen ja laskee sitten sen avulla halutut prosenttijakaumat ja tulostaa ne kolmen desimaalin tarkkuudella työkirjan toiselle työarkille, joka on nimeltään Prosentit. Halutut tunnusluvut tulostuvat kolmannelle työarkille, jonka nimi on Tunnusluvut. Tunnuslukujen lisäksi tulostetaan arvojen lukumäärä.

Tunnuslukujen laskemisessa ohjelma hyödyntää Excelissä jo valmiina olevia funktioita. Esimerkiksi alakvartiili saadaan funktiolla Quartile, jolle annetaan kaksi parametria. Ensimmäisessä parametrissa on muuttujan arvot ja toinen parametri on numero yksi. Yläkvartiili saadaan samalla funktiolla, mutta toiseksi parametriksi annetaan numero kolme. Keskihajontaa ja varianssia ei lasketa, jos muuttuja on vakio. Vakion tapauksessa ei vaihtelua ole ja varianssi on nolla.

Moodin laskemiseen ei ole käytetty valmista funktiota, vaan se on ohjelmoitu itse. Ohjelma määrittää muuttujan suurimman frekvenssin ja etsii sitä vastaavan arvon. Näin saadaan selville muuttujan moodi. Jos moodeja on useita, ohjelma kertoo näistä ensimmäisen ja huomauttaa, että moodeja on enemmän kuin yksi.

Jos käyttäjä valitsee ehdollistavan muuttujan, niin prosenttijakaumat ja tunnusluvut ehdollistetaan tämän muuttujan mukaan. Työarkeille tulostetaan prosentit ehdollistavan muuttujan luokkien mukaan ja koko muuttujan mukaan. Samoin toimitaan tunnuslukuja laskettaessa.

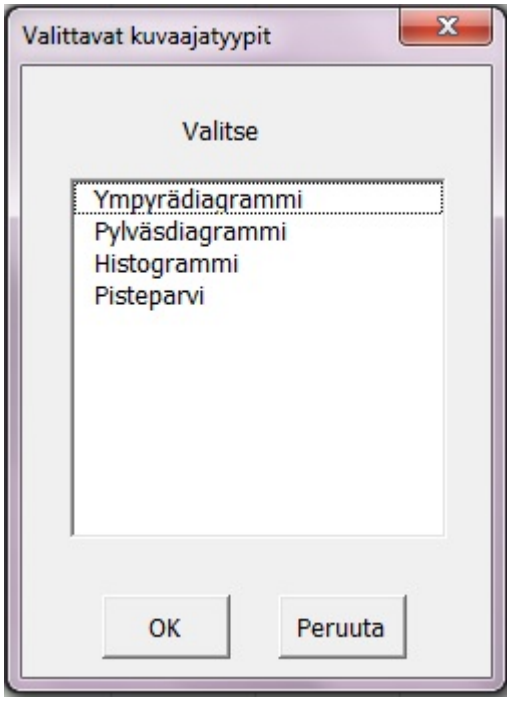


Kuvio 3.3. Tunnusluvut yksiulotteisesta jakaumasta.

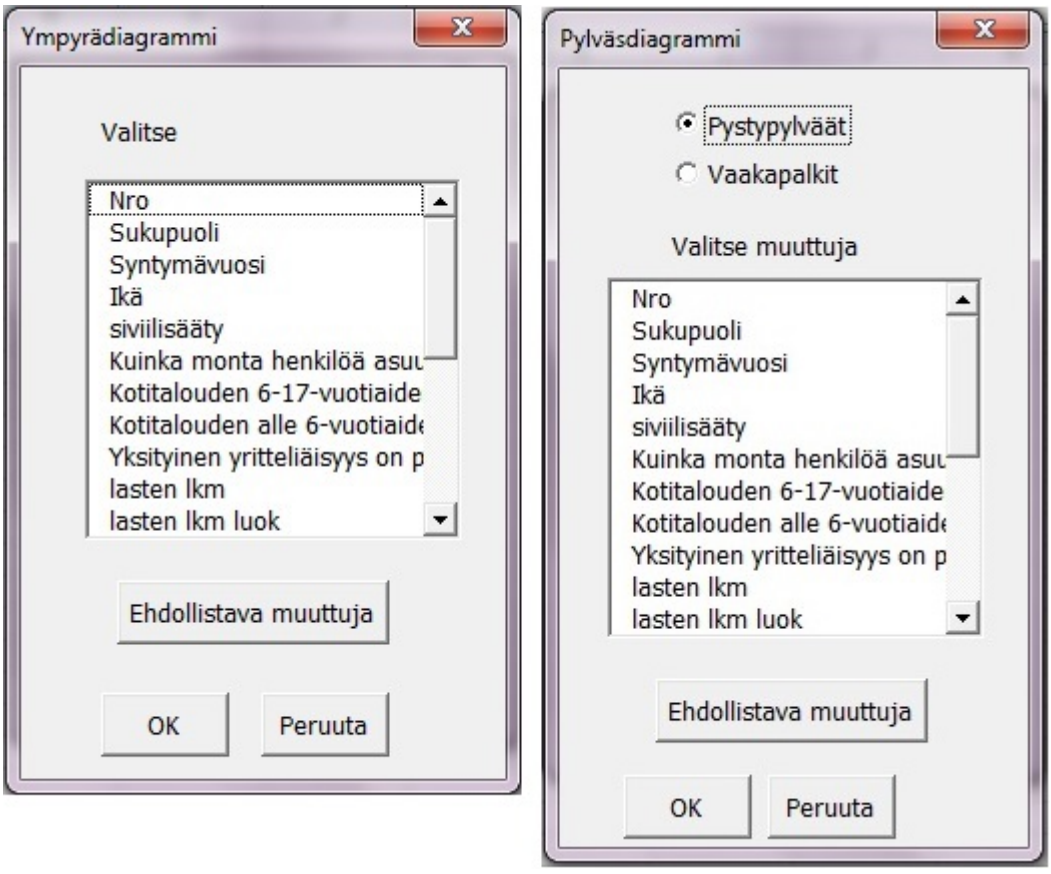
Jos muuttuja sisältää tekstitietoa, niin arvoille annetaan uudet, keinotekoiset numeroarvot alkaen luvusta yksi. Jos käyttäjä haluaa laskea keskihajontaa tai varianssia tällaisesta muuttujasta, ohjelma huomauttaa, että kyseessä on tekstitietoa sisältävä muuttuja, mutta laskee tunnusluvut siitä huolimatta.

3.3 Grafiikka

Pivotin grafiikkavalikosta (kuvio 3.4) voi valita kuvaajaksi ympyrä- tai pylväsdiagrammin, histogrammin tai pisteparven. Näistä vain pisteparvi sopii kahden muuttujan kuvaajaksi. Ohjelma muotoilee kaikkia kuvaajia jonkin verran esimerkiksi siten, että niiden taustat ovat valkoisia.



Kuvio 3.4. Pivotin valikko, jossa voi valita yhden muuttujan kuvaajia.



Kuvio 3.5. Ympyrä- ja pylväsdiagrammin valikot Pivotissa.

3.3.1 Ympyrä- ja pylväsdigrammi

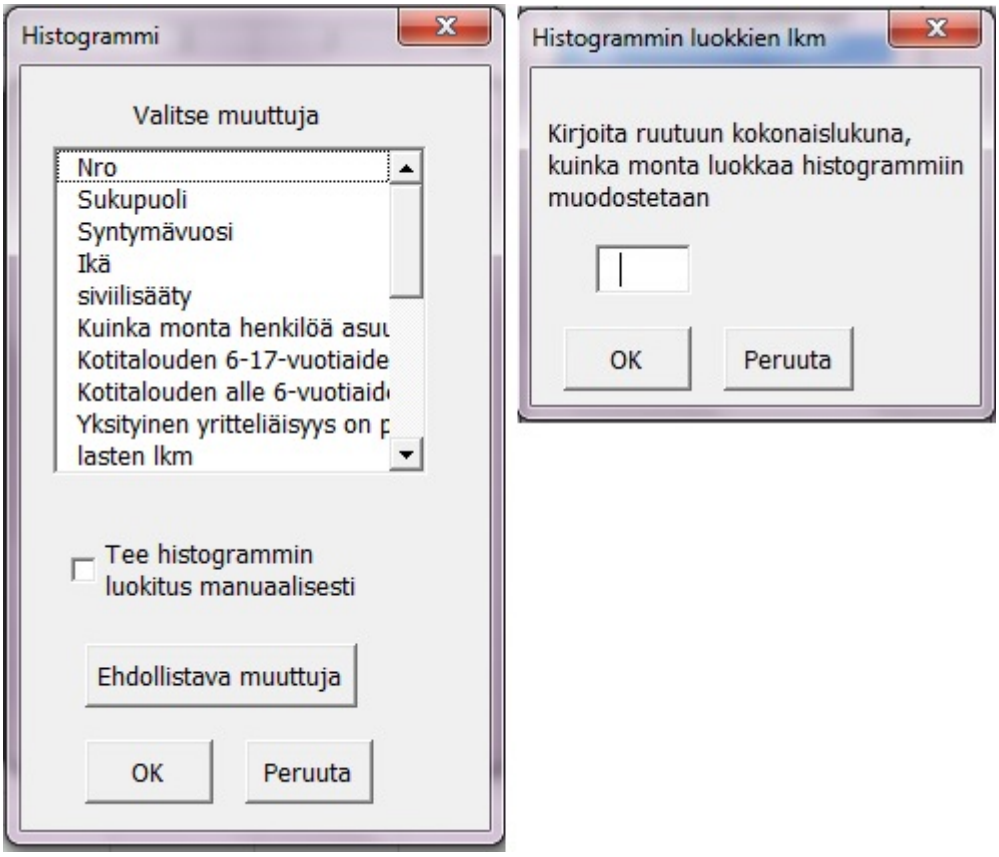
Ympyrädiagrammin valikossa (kuvio 3.5) valitaan muuttuja ja mahdollisesti ehdollistava muuttuja. Uuteen työkirjaan tulostuu ensimmäiselle arkille halutun muuttujan ympyrädiagrammi. Diagrammia on muotoiltu siten, että prosenttiosuudet on lisätty kuvioon yhden desimaalin tarkkuudella. Ohjelma myös otsikoi kuvion muuttujan mukaan. Toiselle työarkille tulostuu PivotTable-taulukko, jossa prosenttiosuudet on esitetty kahden desimaalin tarkkuudella. Jos valitaan ehdollistava muuttuja, kuvioon tulee näkyviin ehdollistavan muuttujan valikko, josta voi valita, halutaanko kuvioon ehdollisen muuttujan kaikki arvot, jokin arvojen yhdistelmä vai yksi arvo. Kuvio muuttuu koko ajan käyttäjän valintojen mukaan. Pylväsdiagrammin tekeminen tapahtuu muuten täysin samalla tavalla kuin ympyrädiagrammin, mutta sen valikossa on mahdollista valita, tehdäänkö pystypylväät vai vaakapalkit. Oletusvalintana on pystypylväät.

3.3.2 Histogrammi

Histogrammin valikko (kuvio 3.6) on ulkoasultaan samanlainen kuin aiemmat grafiikkavalikot. Tämän takana oleva koodi on kuitenkin huomattavasti monimutkaisempi kuin kahden edellisen valikon. Syynä on se, että Excelissä ei ole samalla tavalla valmista kuvaajaa histogrammille kuin esimerkiksi pylväsdiagrammin tapauksessa. Excelissä on kyllä valmiina funktio, jolle annetaan arvot ja luokkarajat ja joka sen jälkeen piirtää pylväsdiagrammin. Kokemattoman käyttäjän on kuitenkin vaikea löytää tämä vaihtoehto, koska se ei löydy samasta paikasta kuin muut kuvaajat, ja se vaatii ensin analyysityökalut-apuohjelman lataamista. Lisäksi kuvaajaa pitäisi funktion jäljiltä vielä muokata. Funktio ei tarkasta sitäkään, onko annettu luokitus tasavälinen. Tätä Excelin valmista työkalua ei ole hyödynnetty Pivotissa, vaan on tehty histogrammin ohjelmointikoodi itse.

Ohjelma on tehty siten, että vain tasavälinen luokitus histogrammiin on mahdollinen. Histogrammin tekeminen aloitetaan luokituksen tekemisestä. Ohjelmaan on tehty valmiiksi kokemattomia käyttäjiä varten eräänlainen yleisluokittelu. Jos muuttuja saa vähemmän kuin neljää erilaista arvoa, mitta-asteikko tuskin on kunnossa ja ohjelma kehottaa käyttäjää tarkistamaan muuttujan mitta-asteikon. Jos eri arvojen lukumäärä on välillä 4–8, ohjelma tekee neljä luokkaa. Välillä 9–12 tehdään viisi luokkaa, välillä 13–16 kuusi luokkaa ja välillä 17–29 kahdeksan luokkaa. Jos eri arvoja on 30 tai enemmän, tehdään 15 luokkaa. Seuraavaksi ohjelma laskee arvojen vaihteluvälin. Luokan pituudeksi tulee vaihteluväli jaettuna luokkien lukumäärällä. Tämän jälkeen määritellään luokkarajat ja lasketaan luokkien frekvenssit. Ohjelma tulostaa uuteen työkirjaan luokkien ylärajat ja frekvenssit ja tekee näiden perusteella pylväsdiagrammin uudelle työarkille. Pylväsdiagrammia muotoillaan niin, että sen pylväät ovat kiinni toisissaan. Tämän jälkeen histogrammi on valmis. Täytyy huomioida, että ohjelma tulostaa x-akselille luokkien ylärajat, ei luokkakeskuksia. Jos käyttäjä ei ole tyytyväinen kuvailtuun yleisluokitteluun, vaan haluaisi erilaisen yleisluokittelun, se on helposti muutettavissa. Koodiin tarvitsee vain vaihtaa käyttäjän haluamat numeroarvot.

On myös mahdollista valita histogrammin luokittelu manuaalisesti. Tällöin ohjelma kysyy käyttäjältä luokkien lukumäärää. Lukumäärän pitää olla välillä 2–100. Jos käyttäjä syöttää jotain muuta, Pivotti kehottaa häntä muuttamaan syötettä. Ohjelma neuvoo syöttämään kokonaislukuja, mutta jos käyttäjä tästä huolimatta syöttää desimaaliluvun, Pivotti pyöristää sen kokonaisluvuksi. Tämän jälkeen histogrammi tehdään samalla tavalla kuin edellä selitettiin. Jos siis jostain syystä halutaan tehdä histogrammi muuttujasta, joka saa vähemmän kuin neljää erilaista arvoa, se onnistuu manuaalisella luokittelulla.



Kuvio 3.6. Pivotin valikko, jossa tehdään histogrammi halutusta muuttujasta sekä manuaalisen luokittelun valikko.

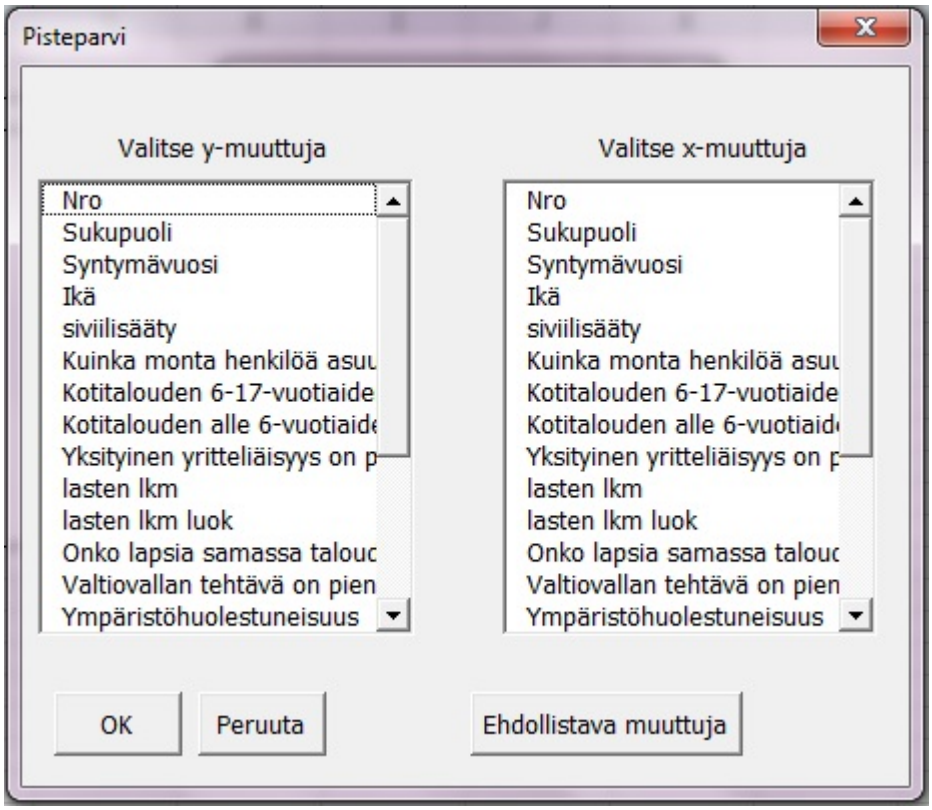
Jos valitaan ehdollistava muuttuja, ohjelma tekee jokaiselle ehdollisen muuttujan luokalle oman histogrammin. Otsikkoon tulostetaan, mikä ehdollisen muuttujan luokista on kyseessä. Näiden lisäksi tehdään koko muuttujan histogrammi. Jos käyttäjä valitsee manuaalisen luokittelun, hän saa määritellä eri luokkien histogrammeille luokkien lukumäärän erikseen. Luokkien lukumäärää siis kysytään yhtä monta kertaa kuin histogrammeja tehdään.

3.3.3 Pisteparvi

Pisteparvea (kuvio 3.7) tehtäessä valitaan muuttujat x- ja y-akseleille. Jos käyttäjä on valinnut saman muuttujan molemmille akseleille, Pivotti huomauttaa tästä ja

piirtää kuvion vasta, kun käyttäjä on valinnut kaksi eri muuttujaa. Pisteparvi otsikoidaan automaattisesti x- ja y-muuttujien mukaan. Excel piirtää pisteparven pisteet salmiakin malliseksi, mutta ohjelma vaihtaa nämä pisteen muotoisiksi. Jos kaikki arvot ovat samassa pisteessä, pisteparvea ei tehdä.

Jos valitaan ehdollistava muuttuja, niin Pivotti tekee jokaiselle ehdollisen muuttujan luokalle oman pisteparven. Otsikkoon tulostetaan, mikä ehdollisen muuttujan luokista on kyseessä. Näiden lisäksi tehdään koko muuttujan pisteparvi.



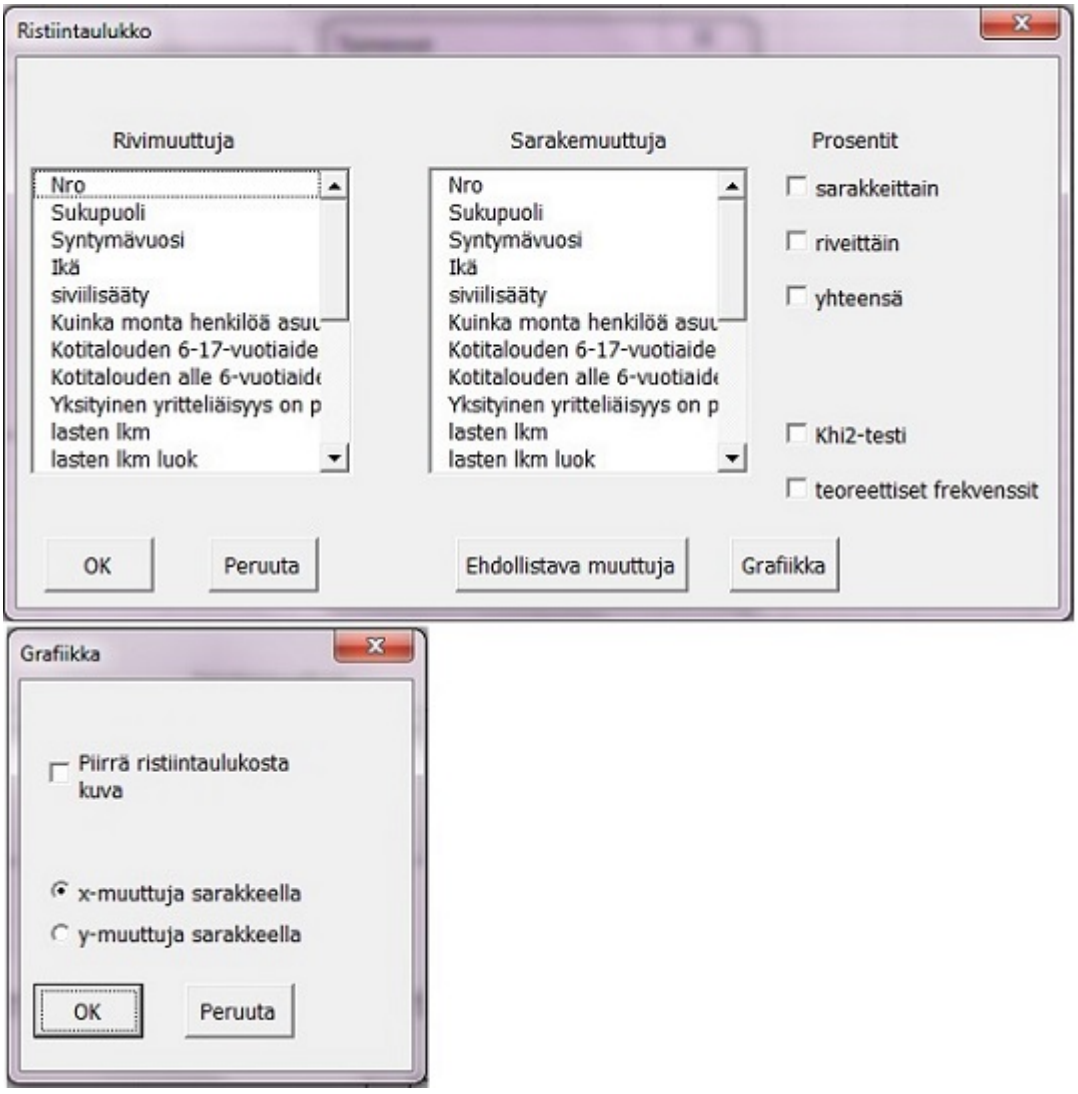
Kuvio 3.7. Pivotin valikko, jossa tehdään pisteparvi halutuista muuttujista.

3.4 Ristiintaulukko ja χ^2-riippumattomuustesti

Kun tehdään ristiintaulukko, valitaan ensin rivi- ja sarakemuuttujat. Tämän jälkeen käyttäjä voi laskea prosentit rivi- tai sarakesuuntaan tai yhteensä. Lisäksi voidaan valita χ^2-testaus ja odotetut frekvenssit. Lomakkeella on myös Grafiikka-painike, jota painamalla saa esille ristiintaulukon kuviota koskevan valikon. Käyttäjä valitsee, onko selittävä muuttuja sarakkeella vai rivillä. Oletusvaihtoehtona on, että selittävä muuttuja on sarakkeella. Jos valitaan vaihtoehto: Piirrä ristiintaulukosta kuva, ohjelma tekee erilliselle työarkille vaakapalkkikuvaajan, jossa selittäjän prosentit summaavat sataan. Kuvio tehdään PivotTable-taulukon perusteella, joten sitä voi muokata samaan tapaan kuin PivotTable-taulukkoakin. Jos kuvioon tehdään muutoksia,

nämä päivittyvät myös PivotTable-taulukkoon. Kuviossa 3.8 esitetään ristiintaulukon ja sen kuvaajan valikot.

Pivotti tekee PivotTable-muotoisen ristiintaulukon valituista muuttujista. Jos χ^2-testaus on valittu, ohjelma tulostaa omalle työarkille testisuureen arvon, käytetyt vapausasteet sekä p-arvon. Lisäksi Pivotti kertoo, ovatko oletukset kunnossa ja ilmoittaa pienimmän odotetun frekvenssin sekä sen, montako prosenttia odotetuista frekvensseistä on < 5. Jos käyttäjä on valinnut laskettavaksi teoreettiset frekvenssit, ne tulostuvat myös samalle työarkille. Jos jokin teoreettisista frekvensseistä saa arvon nolla, χ^2-testisuuretta ei lasketa, koska silloin jakajaan tulisi arvo nolla. Prosentit riveittäin, sarakkeittain ja yhteensä tulostuvat omille arkeilleen. Jos ehtomuuttuja valitaan, kaikki halutut prosentit ja testaukset lasketaan kaikille ehdollistavan muuttujan luokille erikseen.



Kuvio 3.8. Pivotin valikko ristiintaulukon ja χ^2-riippumattomuustestin tekemiselle.

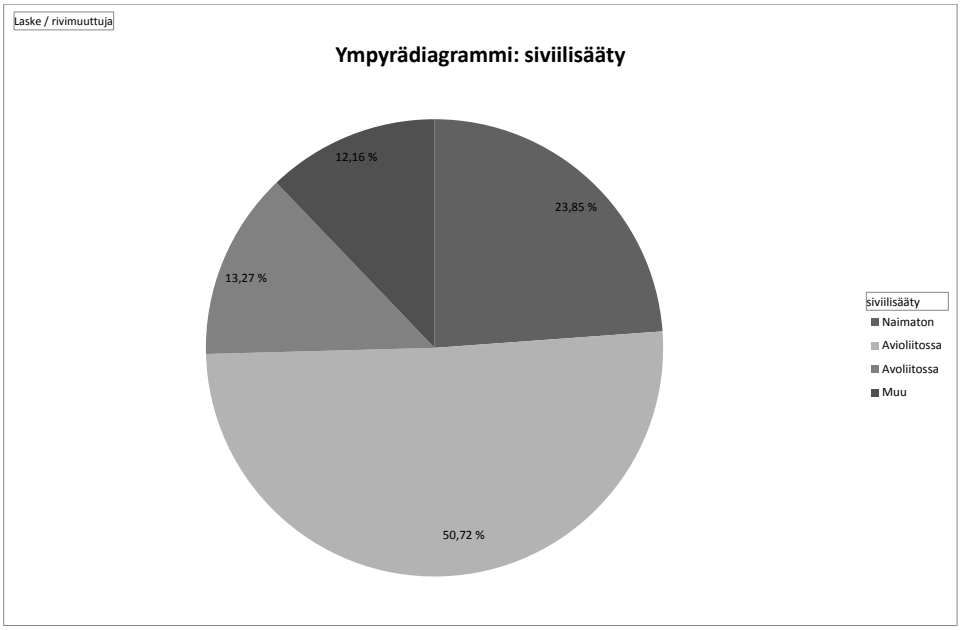
4 Esimerkki Pivotin käytöstä datan analysoinnissa

Esitellään vielä ohjelman käyttöä esimerkkiaineiston tapauksessa. Aineistona on käytetty yleisesti opetuskäytössä ollutta, ympäristömielipiteitä koskevaa aineistoa (kts. alaluku 1.5 sivulla 6). Aineistossa on tilastoyksiköitä 1528 ja muuttujia 115. Tässä esimerkissä tarkastellaan näistä vain muutamia, koska tarkoitus on etupäässä havainnollistaa Pivotin käyttöä ja sen hyödyllisyyttä käytännön tilanteessa. Ohjelman tuottamia kuvia on muotoiltu siten, että niiden tekstejä on suurennettu, jotta lukija näkee luvut paremmin, muuten ne ovat alkuperäisessä muodossaan.

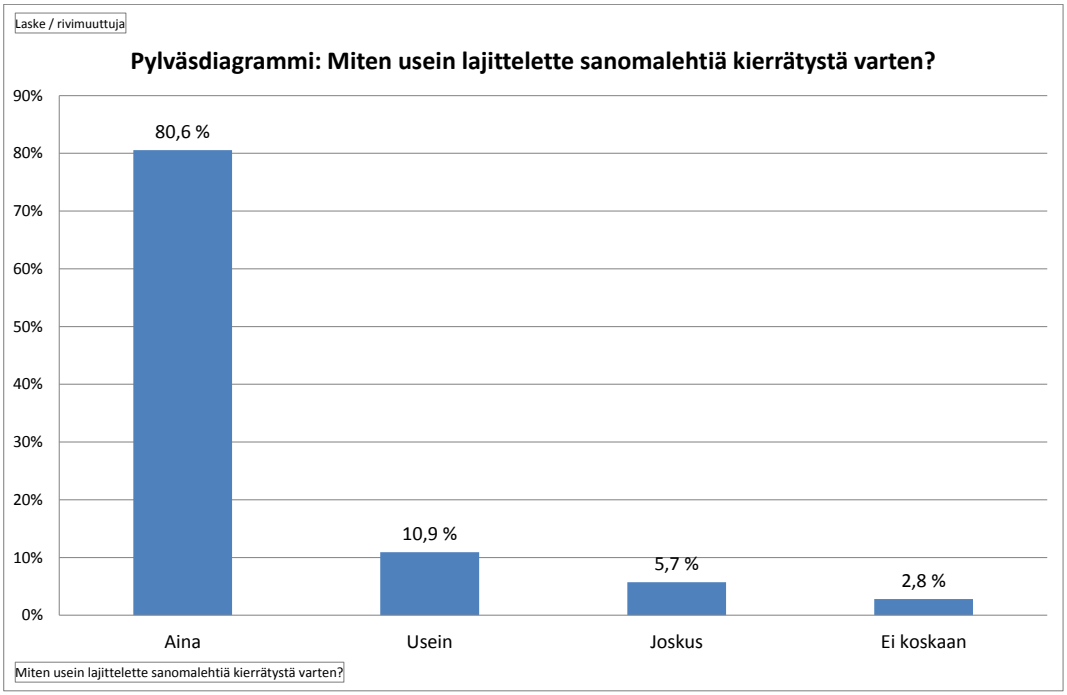
4.1 Aineiston esittely

4.1.1 Ympyrä- ja pylväsdigrammi

Aineistossa naisia on 55 %, yli puolet vastanneista. Tähän kysymykseen vastanneita oli 1523. Kuviossa 4.1 on esitetty ympyrädiagrammi siviilisäädystä. Vähän yli puolet vastaajista on avioliitossa. Vaihtoehtoon muu on yhdistetty vaihtoehdot asumuserossa, eronnut ja leski. Tähän kysymykseen vastasi 1522. Kuviossa 4.2 on esitetty pylväsdiagrammi muuttujasta, joka kertoo, miten usein vastaaja lajittelee sanomalehtiä kierrätystä varten. Yli 80 % vastaajista kertoo lajittelevansa sanomalehdet kierrätystä varten aina ja vain vajaa kolme prosenttia ei lajittele niitä koskaan. Tähän kysymykseen vastanneita oli 1502.



Kuvio 4.1. Ympyrädiagrammi siviilisäädystä.



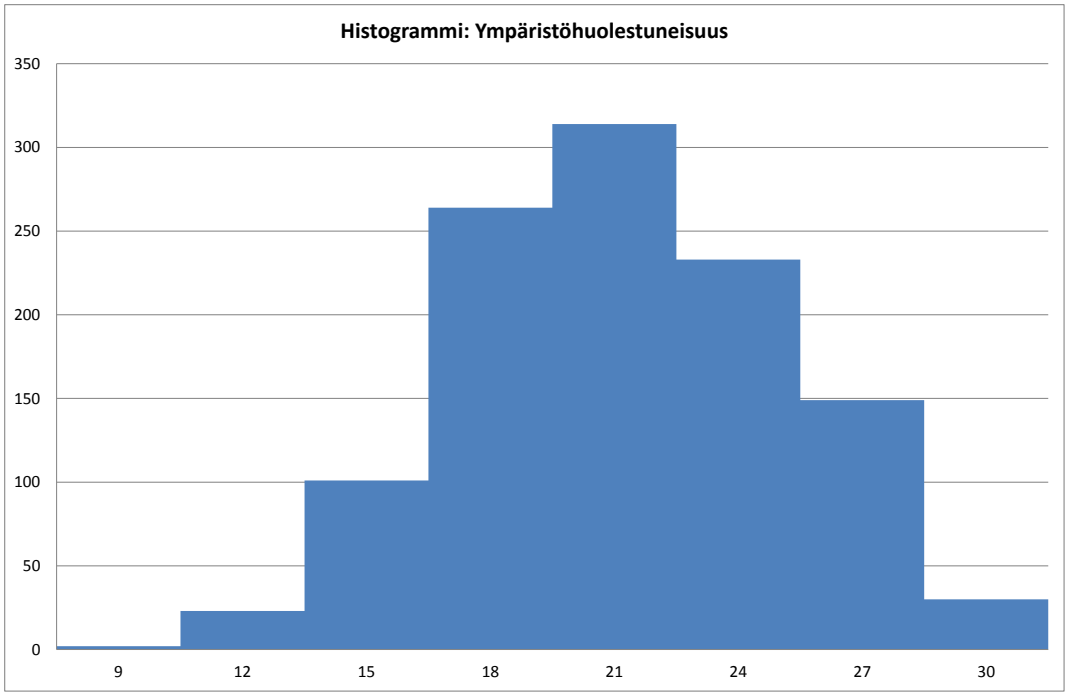
Kuvio 4.2. Pylväsdiagrammi sanomalehtien lajitteluaktiivisuudesta.

4.1.2 Histogrammi automaattisella luokittelulla

Aineistossa on muuttujana kysymys viisi, jossa kysytään kotitalouden 6-17-vuotiaiden henkilöiden lukumäärää. Seuraavana muuttujana on kysymys kuusi, jossa kysytään kotitalouden alle 6-vuotiaiden lasten lukumäärää. Näistä kahdesta muuttujasta on muodostettu uusi muuttuja, lasten lukumäärä samassa taloudessa, summaamalla muuttujat viisi ja kuusi. Tästä uudesta muuttujasta on edelleen tehty uusi muuttuja, onko alaikäisiä lapsia samassa taloudessa. Jos lasten lukumäärä samassa taloudessa on nolla, saa uusi muuttuja arvon "ei"ja muussa tapauksessa arvon "kyllä". Kaikki vastaajat olivat vastanneet kysymyksiin viisi ja kuusi. Talouksia, joissa ei ole alaikäisiä lapsia on 64 %.

Kysymyksessä 19 tiedusteltiin, kuinka vaaralliseksi ympäristölle vastaaja kokee seuraavat asiat: teollisuuden aiheuttama ilmansaastuminen, maanviljelyksessä käytettävät hyönteismyrkyt ja kemikaalit, Suomen jokien, järvien ja vesistöjen saastuminen, kasvihuoneilmiön aiheuttama maapallon lämpötilan kohoaminen, joidenkin viljalajien geneettinen muuntelu ja ydinvoimalat. Näihin vastausvaihtoehtoina oli 1 = äärimmäisen vaarallista, 2 = hyvin vaarallista, 3 = melko vaarallista, 4 = ei kovin vaarallista, 5 = ei lainkaan vaarallista ja 6 = en osaa sanoa. Näistä muuttujista on muodostettu uusi muuttuja ympäristöhuolestuneisuus siten, että en osaa sanoa arvot muutettiin puuttuvaksi tiedoksi ja tämän jälkeen asteikot käännettiin niin, että 5 = äärimmäisen vaarallista ja 1 = ei lainkaan vaarallista. Tämän jälkeen kysymyksen 19 kaikki kuusi vaihtoehtoa summattiin summamuuttujaksi. Jos vastaaja oli jättänyt johonkin näistä kysymysvaihtoehdoista vastaamatta, summamuuttujaa ei laskettu. Ky-

seinen summamuuttuja kertoo siitä, kuinka vaarallisena ympäristölle henkilö pitää edellä mainittuja seikkoja. Mitä suurempi luku, sitä vaarallisempana vastaaja näitä pitää. Summamuuttuja voi saada arvoja 6–30. Kuviossa 4.3 on esitetty ympäristöhuolestuneisuuden histogrammi, joka on toteutettu Pivotin automaattisella luokittelulla. Nähdään, että jakauma on lähes symmetrinen. Muuttujan keskiarvo on 20,5 ja mediaani 20,0.

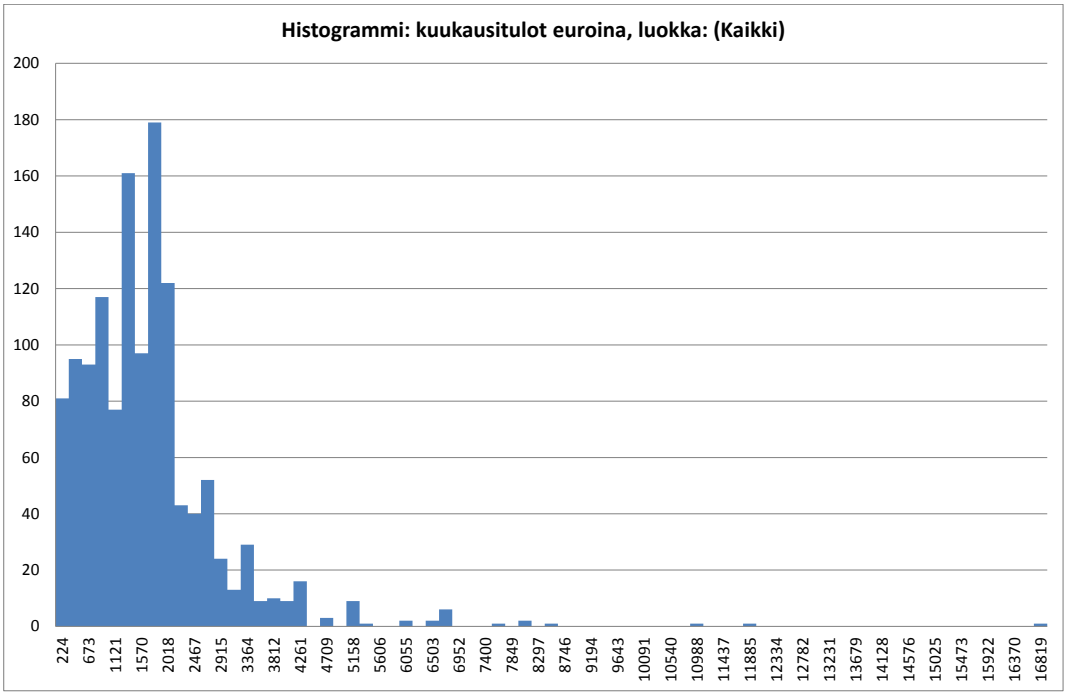


Kuvio 4.3. Pivotilla tehty histogrammi ympäristöhuolestuneisuudesta.

4.1.3 Histogrammi manuaalisella luokittelulla ja ehdollistettuna

Tarkastellaan seuraavaksi kuukausitulojen jakaumaa sukupuolen mukaan. Aineistossa kuukausitulot on kerrottu markkoina. Tehdään ensin uusi muuttuja, kuukausitulot euroina, jakamalla tulot markkoina luvulla 5,94573. Piirretään histogrammi tästä uudesta muuttujasta ja valitaan ehdollistavaksi muuttujaksi sukupuoli. Koska kuukausitulojen jakauma on oikealle vino, automaattisen histogrammin teon 30 luokkaa eivät riitä kuvaamaan jakaumaa hyvin, koska melkein kaikki arvot sijoittuvat ensimmäisiin luokkiin (liite B.1). Valitaan manuaalinen luokittelu ja annetaan luokkien lukumääräksi tarpeeksi iso arvo. Tässä tilanteessa kaikille kolmelle ohjelman tekemälle kuviolle on annettu luokkien lukumääräksi 75. Kuviossa 4.4 on kuukausitulojen histogrammi. Nähdään, että jakauma on oikealle vino ja poikkeavan suuria arvoja on muutamia. Miesten ja naisten kuukausitulojen histogrammit löytyvät liitteistä (B.2 ja B.3). Tulkinnassa on huomattava, että x-akselilla näkyvät arvot ovat luokkien ylärajoja, eivät luokkakeskuksia. Miesten ja naisten jakaumissa

on eroja siten, että miesten tulot ovat suurempia. Molemmat jakaumat ovat kuitenkin oikealle vinoja.



Kuvio 4.4. Pivotilla tehty histogrammi kuukausituloista.

4.1.4 Tunnusluvut

Lasketaan vielä tunnuslukuja kuukausituloille sukupuolen mukaan (taulukko 4.1). Voidaan havaita, että poikkeavat arvot vaikuttavat keskiarvoon varsinkin miehillä. Miesten kaikki arvot ovat suurempia kuin naisten, kvartiileissakin miesten tunnusluvut ovat satoja euroja suurempia kuin naisten. Näyttäisi siis siltä, että miesten kuukausitulot ovat suurempia kuin naisten. Täytyy kuitenkin muistaa, että taustalla on muitakin asiaan vaikuttavia seikkoja kuin sukupuoli.

4.2 Pisteparvi

Tarkastellaan kuukausitulojen ja ympäristöhuolestuneisuuden riippuvuutta sukupuolittain. Piirretään muuttujista pisteparvi ja valitaan ehdollistavaksi muuttujaksi sukupuoli. Ohjelma tekee kolme eri kuviota. Kuviossa 4.5 on koko aineiston pisteparvi. Nähdään, että kuukausituloilla ja ympäristöhuolestuneisuudella ei näyttäisi olevan riippuvuutta. Miesten ja naisten pisteparvissa (liitteet B.4 ja B.5) ei ole suuria eroja, eikä kummassakaan näyttäisi olevan riippuvuutta. Miehillä on yksi poikkeavan suuri arvo kuukausituloissa ja naisilla yksi pieni (alle kymmenen) arvo ympäristöhuolestuneisuudessa. Tulot eivät siis kuvaajan perusteella vaikuta siihen, miten huolissaan ihminen on ympäristöstä.

Taulukko 4.1. Kuukausitulojen tunnuslukuja sukupuolen mukaan

Kuukausitulot euroina	Mies	Nainen	Kaikki
keskiarvo	1814,6	1363,3	1569,3
mediaani	1681,9	1335,8	1429,6
moodi	1681,9	1681,9	1681,9
keskihajonta	1432,6	1034,6	1252,8
varianssi	2052308,5	1070370,7	1569434,6
minimi	0,0	0,0	0,0
maksimi	16818,8	10932,2	16818,8
alakvartiili	1009,1	672,8	767,9
yläkvartiili	2186,4	1681,9	2018,3
vaihteluväli	16818,8	10932,2	16818,8
prosentuaali: 0,1	406,4	336,4	336,4
prosentuaali: 0,9	3363,8	2354,6	2859,2
lkm	594	702	1296

4.3 Ristiintaulukko ja χ^2-riippumattomuustesti

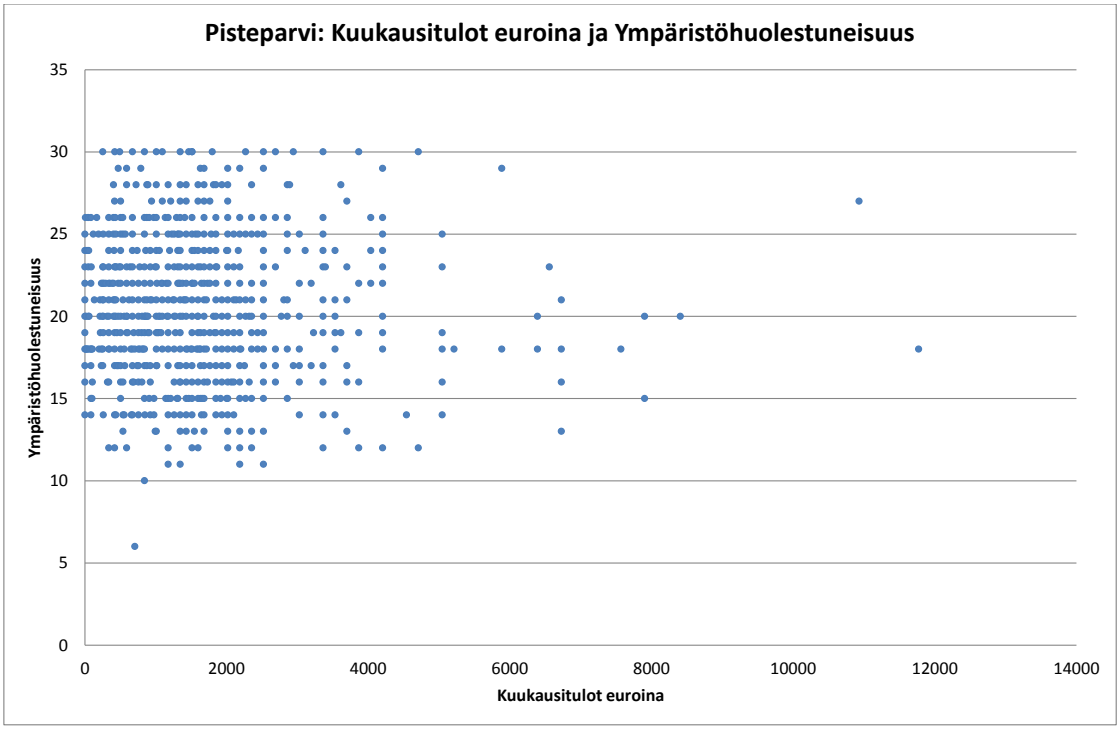
Havainnollistetaan ohjelman antamia tulosteita, kun tehdään ristiintaulukko. Edellisessä esimerkissä tarkasteltiin tulojen vaikutusta ympäristöön liittyviin asenteisiin. Tarkastellaan seuraavaksi, vaikuttavatko tulot sanomalehtien lajitteluaktiivisuuteen. Sanomalehtien lajittelua voisi pitää tekona, joka hyödyttää ympäristöä. Tutkitaan asiaa vielä miehillä ja naisilla erikseen. Kuukausitulot muuttuja on luokiteltu kahteen luokkaan likimain mediaanin kohdalta. Nyt hypoteeseina ovat:

H_0: kuukausituloilla ja sanomalehtien lajittelulla ei ole riippuvuutta
H_1: kuukausituloilla ja sanomalehtien lajittelulla on riippuvuutta.

Valitaan ristiintaulukon valikosta oikeat muuttujat, prosentit sarakkeittain sekä χ^2-testi. Tämän jälkeen ohjelma tulostaa seuraavat tiedot sekä prosentit sarakkeittain ja vielä lisäksi absoluuttiset frekvenssit.

χ^2: 7,73
df: 3
p-arvo: 0,052
Oletukset ovat kunnossa.
Pienin teoreettinen frekvenssi: 17,12
Montako prosenttia teoreettisista frekvensseistä < 5: 0

Miehet
χ^2: 17,05
df: 3
p-arvo: 0,00069



Kuvio 4.5. Pivotilla tehty pisteparvi kuukausituloista ja ympäristöhuolestuneisuudesta.

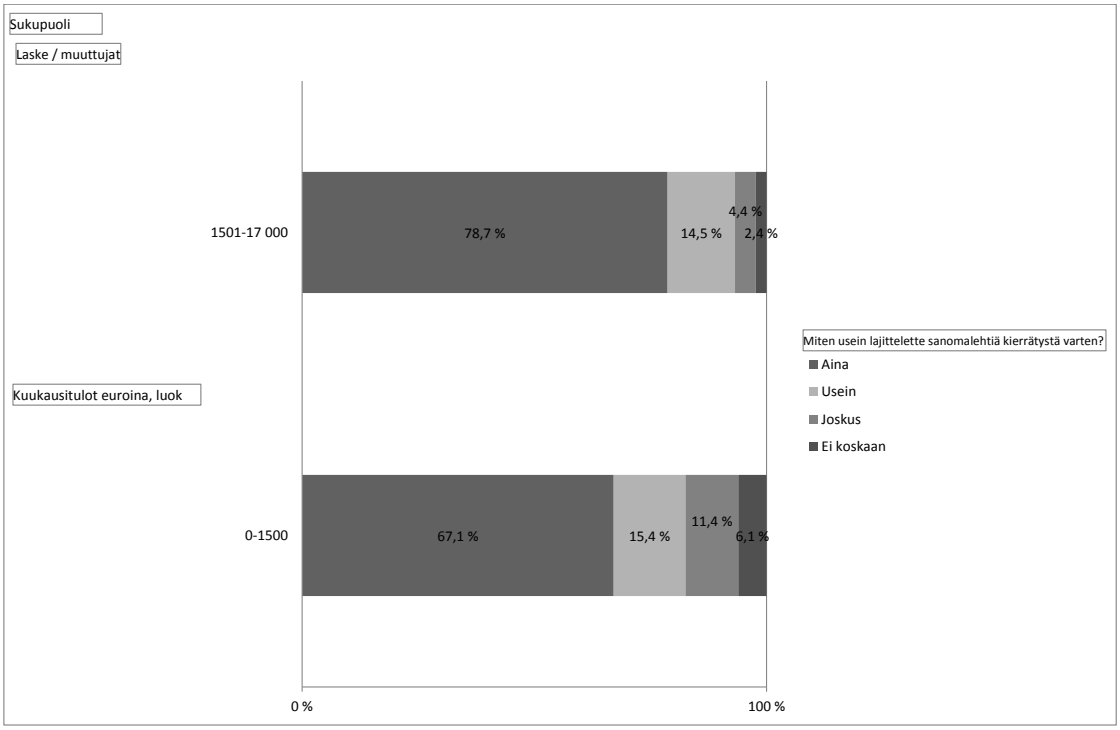
Oletukset ovat kunnossa.
Pienin teoreettinen frekvenssi: 9,69
Montako prosenttia teoreettisista frekvensseistä < 5: 0

Naiset
χ^2: 1,33
df: 3
p-arvo: 0,721
Oletukset ovat kunnossa.
Pienin teoreettinen frekvenssi: 4,96
Montako prosenttia teoreettisista frekvensseistä < 5: 12,5

Oletukset ovat kunnossa kaikissa tapauksissa. Kaksi p-arvoista on $> 0,01$ ja miesten tapauksessa p-arvo $< 0,01$. Yleisesti kuukausitulot eivät siis vaikuta sanomalehtien lajitteluaktiivisuuteen. Tällöin H_0 hyväksytään. Miesten tapauksessa H_0 kuitenkin hylätään ja päätellään, että riippuvuutta on. Kun tarkastellaan miesten ehdollisia prosenttijakaumia (taulukko 4.2), näyttää siltä, että miehet, joilla on isommat kuukausitulot lajittelevat sanomalehtiä vähän useammin kuin pienempituloiset. Sama asia nähdään myös ohjelman piirtämästä kuvasta (kuvio 4.6).

Taulukko 4.2. Ristiintaulukon prosentit sarakkeittain.

Sanomalehtien lajittelu	0–1500	1501–17 000
Aina	0,865	0,892
Usein	0,079	0,059
Joskus	0,039	0,031
Ei koskaan	0,017	0,017



Kuvio 4.6. Pivotilla tehty kuvio ristiintaulukosta, jossa tarkastellaan tulojen vaikutusta sanomalehtien lajitteluaktiivisuuteen miesten tapauksessa.

5 Yhteenveto

Tutkielmasta on käynyt ilmi, että Pivotti auttaa kokematonta käyttäjää hyödyntämään Excelin PivotTablea ja nopeuttaa taulukoiden ja kuvioiden muodostusta. Pivotti on siis hyödyllinen aineiston kuvailemiseen ja analysoimiseen. Se sisältää neljä eri kuvaajatyyppiä: ympyrä- ja pylväsdiagrammin, histogrammin ja pisteparven. Histogrammista voi valita automaattisen luokittelun, jolloin ohjelma tekee luokittelun tai käyttäjä voi tehdä sen manuaalisesti. Yksiulotteisesta muuttujasta Pivotilla voi laskea prosenttiosuudet ja tunnusluvut. Ohjelmalla pystyy tekemään myös ristiintaulukon, piirtämään siitä kuvaajan sekä laskemaan χ^2-riippumattomuustestin ja teoreettiset frekvenssit. Kaikkiin tarkasteluihin voi lisätä ehdollistavan muuttujan. Pivotti on laajennettavissa käsittämään muitakin tilastollisia menetelmiä.

Ohjelmasta on tehty mahdollisimman helppokäyttöinen ja se sisältää useita tarkistuksia. Tarkistuksia tarvitaan, jotta käyttäjä saisi oikeanlaiset virheilmoitukset sen sijaan, että ohjelman suoritus vain keskeytyisi. Ohjelma esimerkiksi antaa virheilmoituksen, jos yritetään laskea tunnuslukuja tekstitietoisesta muuttujasta. Luvussa kolme käydään läpi näitä tarkistuksia ja se toimii manuaalina Pivotin käyttöä varten. Siinä ohjeistetaan yksityiskohtaisesti ja kuvien avulla, miten Pivottia käytetään.

Luvussa neljä osoitetaan käytännössä, miten Pivottia käytetään esimerkkiaineiston analysoimisessa. Luvussa käydään läpi kaikki kuvaajatyypit, joita Pivotilla pystyy tekemään sekä lasketaan tunnuslukuja. Myös ristiintaulukosta, siitä tehtävästä kuvaajasta ja χ^2-riippumattomuustestistä on esimerkki. Lisäksi havainnollistetaan ehtomuuttujan käyttöä. Kaikki saadut tulokset on myös tulkittu.

Lähteet

Leppälä, R. (2012), "Tilastollisten menetelmien perusteet II", luentorunko, Tampereen yliopisto, informaatiotieteiden yksikkö. Saatavilla Internetistä: http://www.sis.uta.fi/tilasto/tiltp3/kevat2012/luennot.pdf.

Manninen, P. (2006), "VBA, Visual Basic for Application (Excel)", luentomoniste, Tampereen yliopisto, informaatiotieteiden yksikkö.

International social survey programme: ympäristö II, 2000: Suomen aineisto [elektroninen aineisto]. FSD0115, versio 2.1 (2006-05-08). Tampere: Tampereen yliopisto. Sosiologian ja sosiaalipsykologian laitos & Tampere: Yhteiskuntatieteellinen tietoarkisto & Helsinki: Tilastokeskus [tuottajat], 2000. Tampere: Yhteiskuntatieteellinen tietoarkisto [jakaja], 2006.

Liite A

Liite: Kyselylomake

KYSELYLOMAKE

Tämä kyselylomake on osa Yhteiskuntatieteelliseen tietoarkistoon arkistoitua tutkimusaineistoa

FSD0115 ISSP 2000 : ympäristö II : Suomen aineisto

Kyselylomaketta hyödyntävien tulee viitata siihen asianmukaisesti lähdeviitteellä.

Lisätiedot: http://www.fsd.uta.fi/

QUESTIONNAIRE

This questionnaire is part of the following dataset, archived at the Finnish Social Science Data Archive:

FSD0115 ISSP 2000 : Environment II : Finnish Data

If this questionnaire is used or referred to in any publication, the source must be acknowledged by means of an appropriate bibliographic citation.

More information: http://www.fsd.uta.fi/

Aluksi muutama Teitä ja kotitalouttanne koskeva kysymys:
Rengastakaa mielestänne sopivinta vaihtoehtoa vastaava numero tai kirjoittakaa vastauksenne sille varattuun tilaan.

1. Sukupuolenne?

mies 1
nainen 2

2. Syntymävuotenne? 19 _____

3. Oletteko tällä hetkellä…

naimaton 1
avioliitossa 2
avoliitossa 3
asumuserossa..... 4
eronnut 5
leski 6
muu 7

4. Kuinka monta henkilöä asuu samassa kotitaloudessa Teidän kanssanne? _____ henkilöä
Laskekaa itsenne mukaan.

5. Kotitaloutenne *6-17-vuotiaiden henkilöiden* **lukumäärä?** _____ 6-17-vuotiasta

6. Kotitaloutenne *alle 6-vuotiaiden lasten* **lukumäärä?** _____ alle 6-vuotiasta

7. Asutteko?

kaupungin keskustassa 1
esikaupunkialueella tai kaupunkilähiössä..... 2
kuntakeskuksessa tai muussa taajamassa .. 3
maaseudun haja-asutusalueella 4

Seuraavaksi muutama yhteiskuntaa yleisesti koskeva kysymys:

8. Mitä mieltä olette seuraavista väittämistä?
Rengastakaa jokaiselta riviltä näkemystänne parhaiten vastaava vaihtoehto.

		Täysin samaa mieltä	Samaa mieltä	En ole samaa mieltä enkä eri mieltä	Eri mieltä	Täysin eri mieltä	*En osaa sanoa*
a)	Yksityinen yritteliäisyys on paras tapa ratkaista Suomen taloudelliset ongelmat	1	2	3	4	5	*6*
b)	Valtiovallan tehtävä on pienentää tuloeroja suuri ja pienituloisten välillä	1	2	3	4	5	*6*

17. Kuinka paikkansa pitäviä seuraavat väitteet mielestänne ovat?
Rengastakaa jokaiselta riviltä näkemystänne parhaiten vastaava vaihtoehto.

		On ehdotto-masti totta	On luulta-vasti totta	Ei ole totta	Ei missään tapauksessa ole totta	*En osaa sanoa*
a)	Antibiootit tehoavat vain bakteereihin mutta eivät viruksiin	1	2	3	4	*5*
b)	Ihmiset ovat kehittyneet varhaisemmista eläinlajeista	1	2	3	4	*5*
c)	Kaikki kemikaalit voivat aiheuttaa syöpää, jos niitä nauttii tarpeeksi paljon	1	2	3	4	*5*
d)	Vähäisenkin radioaktiivisuuden saaminen aiheuttaa varman kuoleman	1	2	3	4	*5*
e)	Kasvihuoneilmiön aiheuttaja on maan ilmakehässä oleva aukko	1	2	3	4	*5*
f)	Joka kerta kun käytämme hiiltä, öljyä tai kaasua, kasvihuoneilmiö pahenee	1	2	3	4	*5*

18. Kuinka vaarallisia ovat mielestänne seuraavat asiat?
Rengastakaa jokaiselta riviltä näkemystänne parhaiten vastaava vaihtoehto.

		Äärimmäisen vaarallisia	Hyvin vaarallisia	Melko vaarallisia	Ei kovin vaarallisia	Ei lainkaan vaarallisia	*En osaa sanoa*
a)	Autojen aiheuttama ilmansaastuminen *yleisesti ympäristölle*	1	2	3	4	5	*6*
b)	Autojen aiheuttama ilmansaastuminen *Teille tai perheellenne*	1	2	3	4	5	*6*

19. Yleisesti ottaen kuinka vaarallisia *ympäristölle* **ovat mielestänne seuraavat asiat?**
Rengastakaa jokaiselta riviltä näkemystänne parhaiten vastaava vaihtoehto.

		Äärimmäisen vaarallisia	Hyvin vaarallisia	Melko vaarallisia	Ei kovin vaarallisia	Ei lainkaan vaarallisia	*En osaa sanoa*
a)	Teollisuuden aiheuttama ilmansaastuminen	1	2	3	4	5	*6*
b)	Maanviljelyksessä käytettävät hyönteismyrkyt ja kemikaalit	1	2	3	4	5	*6*
c)	Suomen jokien, järvien ja vesistöjen saastuminen	1	2	3	4	5	*6*
d)	Kasvihuoneilmiön aiheuttama maapallon lämpötilan kohoaminen	1	2	3	4	5	*6*
e)	Joidenkin viljalajien geneettinen muuntelu	1	2	3	4	5	*6*
f)	Ydinvoimalat	1	2	3	4	5	*6*

29. Rengastakaa vaihtoehto, joka on lähinnä käsitystänne.

		Erittäin paljon	Paljon	En paljon enkä vähän	Vähän	En lain-kaan	*En osaa sanoa*
a)	Luotatteko ympäristöuhkia koskevissa kysymyksissä asiantuntijoihin?	1	2	3	4	5	*6*
b)	Vaikuttavatko asiantuntijoiden kannanotot toimintaanne ympäristöasioissa?.....	1	2	3	4	5	*6*

30. Kuinka usein lajitellette seuraavia jätteitä kierrätystä varten?
Rengastakaa joka riviltä näkemystänne parhaiten vastaava vaihtoehto.

		Aina	Usein	Joskus	En kos-kaan
a)	Lasia	1	2	3	4
b)	Tölkkejä	1	2	3	4
c)	Sanomalehtiä.....	1	2	3	4

31. Miten usein rajoitatte omaa auton käyttöänne ympäristösyistä?
Rengastakaa näkemystänne parhaiten vastaava vaihtoehto.

Minulla ei ole autoa tai ajokorttia 0

aina 1
usein 2
joskus..... 3
en koskaan 4

32. Oletteko jäsenenä jossakin ryhmässä, jonka tärkeimpänä tavoitteena on luonnon- tai ympäristönsuojelu?

kyllä..... 1
en 2

33. Oletteko viimeksi kuluneen viiden vuoden aikana…
Rengastakaa jokaiselta riviltä näkemystänne parhaiten vastaava vaihtoehto.

		Kyllä	En
a)	allekirjoittanut jotakin ympäristönsuojelua koskeneen adressin tai vastaavan?	1	2
b)	lahjoittanut rahaa jollekin ympäristönsuojelujärjestölle tai –ryhmälle?.....	1	2
c)	ottanut osaa mielenosoitukseen tai marssiin jonkun ympäristöasian vuoksi?.....	1	2

Seuraavat kysymykset koskevat kaikkia vastaajia *(myös 15 – 17-vuotiaita).*

55. Jos eduskuntavaalit järjestettäisiin nyt, niin minkä puolueen tai muun ryhmittymän ehdokasta äänestäisit?
Rengastakaa sopivin vaihtoehto.

Suomen sosiaalidemokraattinen puolue (SDP)	1
Suomen keskusta (Kesk)	2
Kansallinen kokoomus (Kok)	3
Vasemmistoliitto (Vas)	4
Ruotsalainen kansanpuolue (RKP)	5
Vihreä liitto (Vihr)	6
Suomen kristillinen liitto (SKL)	7
Perussuomalaiset (PS)	8
Remonttiryhmä (Rem)	9
Jokin muu puolue tai ryhmittymä	10
En äänestäisi	*11*
En osaa sanoa	*12*
En halua sanoa	*13*

56. Politiikasta keskusteltaessa puhutaan usein vasemmistosta ja oikeistosta. Mihin kohtaan sijoittaisitte oman kantanne tällä asteikolla yleensä ottaen?
Rengastakaa näkemystänne parhaiten vastaava vaihtoehto.

Vasemmisto	1	2	3	4	5	6	7	8	9	10	Oikeisto

57. Kuulutteko kirkkoon tai uskonnolliseen yhteisöön? Jos kuulutte niin mihin?
Rengastakaa sopivin vaihtoehto.

evankelis-luterilaiseen kirkkoon	1
ortodoksiseen kirkkoon	2
muuhun kristilliseen kirkkoon tai yhteisöön	3
muuhun uskonnolliseen yhteisöön	4
En kuulu kirkkoon tai muuhun uskonnolliseen yhteisöön	*5*

58. Usein puhutaan *sosiaaliluokista tai sosiaaliryhmistä*. Mihin sosiaaliluokkaan tai -ryhmään katsotte itse kuuluvanne?
Rengastakaa sopivin vaihtoehto.

alaluokkaan	1
työväenluokkaan	2
alempaan keskiluokkaan	3
ylempään keskiluokkaan	4
yläluokkaan	5
En osaa sanoa	*6*
Ei mihinkään näistä luokista	*7*

59. Kuinka suuret ovat keskimääräiset kuukausitulonne *veroja vähentämättä* **(= bruttotulot)?**

_____ markkaa kuukaudessa

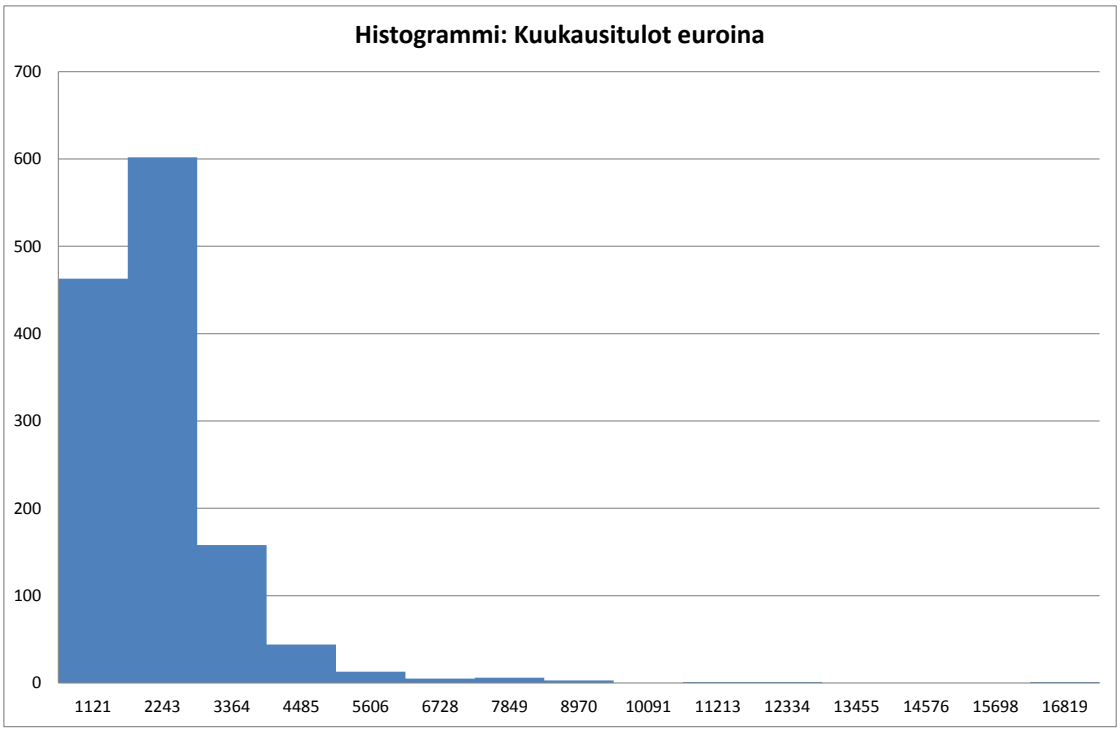
60. Kuinka suuret ovat keskimäärin kotitaloutenne yhteenlasketut kuukausitulot *veroja vähentämättä* **(= bruttotulot) mukaan laskien veronalaiset sosiaalietuudet?**

_____ markkaa kuukaudessa

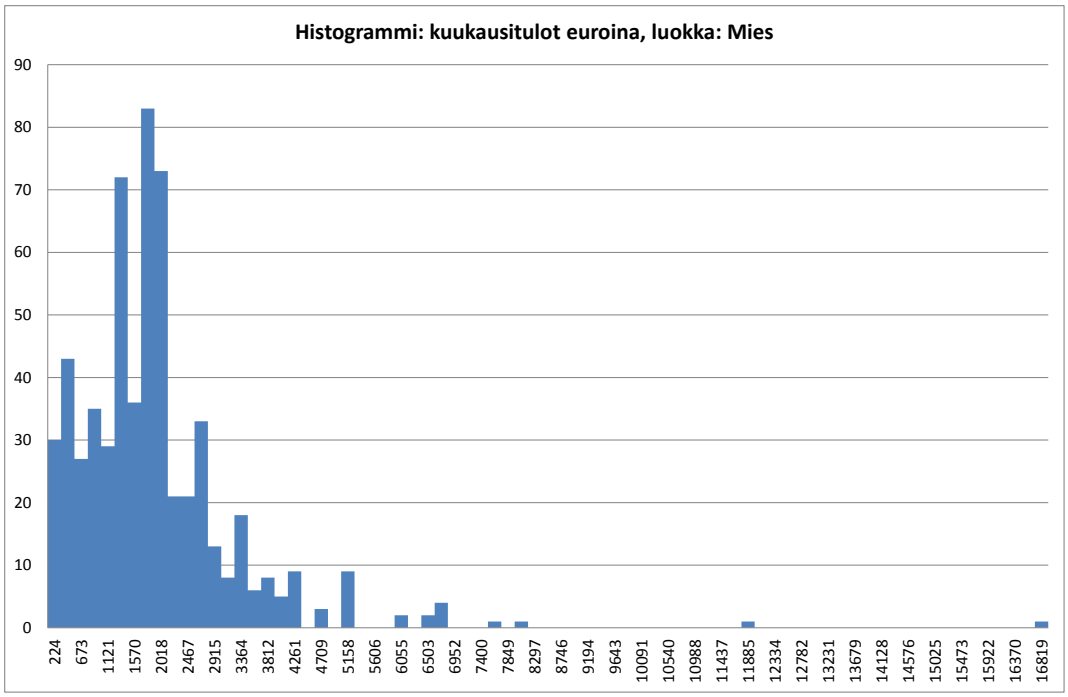
KIITOKSIA VAIVANNÄÖSTÄNNE!

Liite B

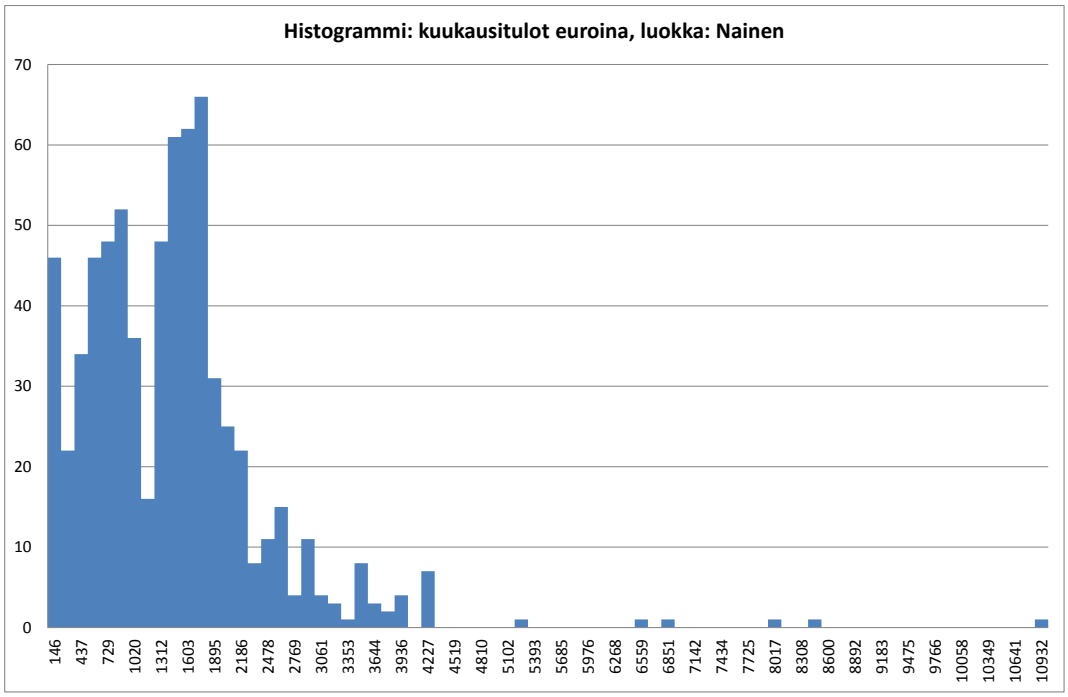
Liite: Kuvaajia esimerkkiaineistosta



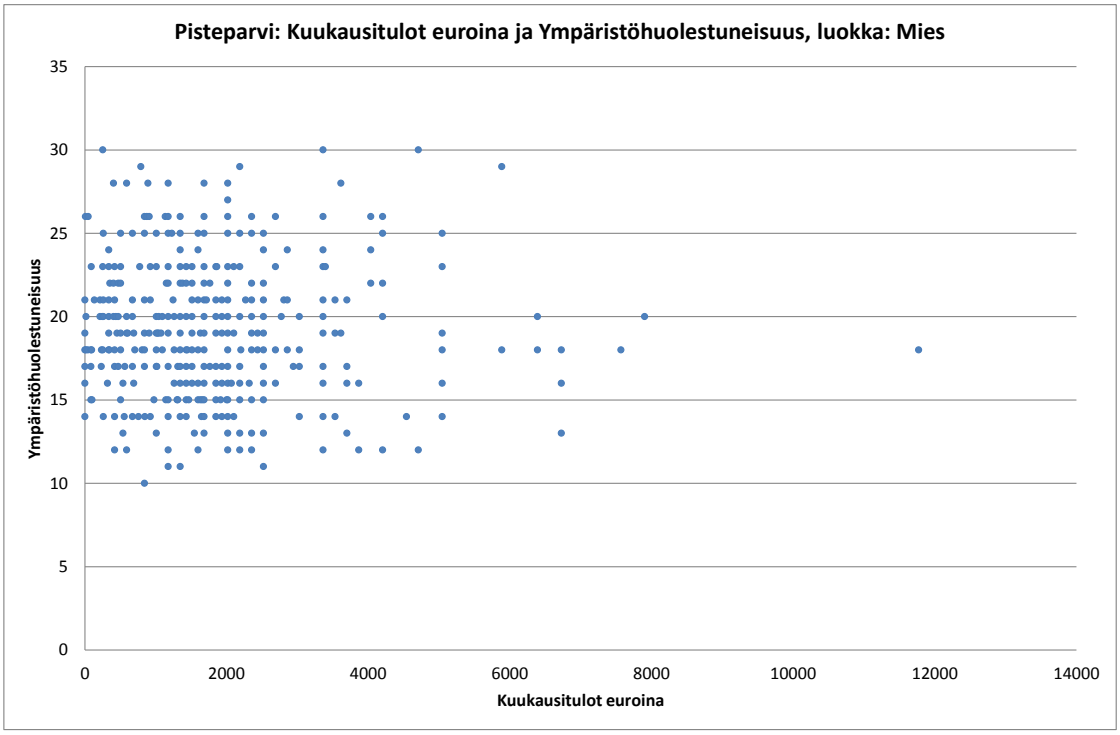
Kuva B.1. Pivotilla tehty histogrammi kuukausituloista, kun käytetään automaattista luokittelua.



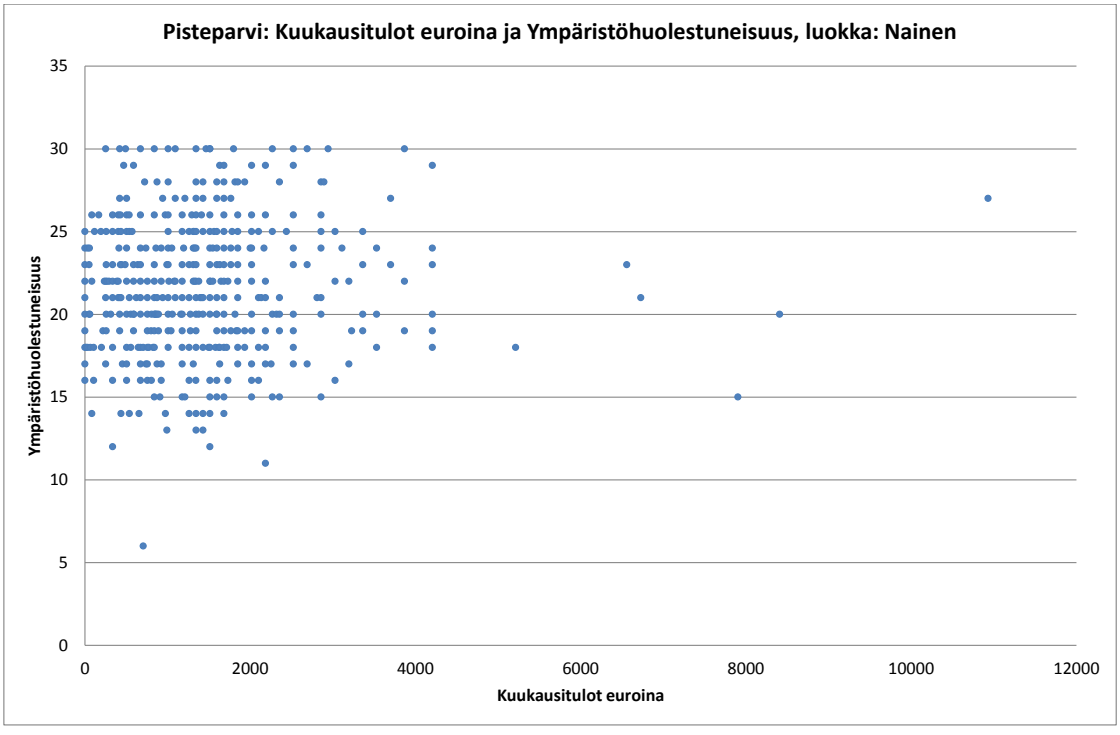
Kuva B.2. Pivotilla tehty histogrammi miesten kuukausituloista manuaalisella luokittelulla.



Kuva B.3. Pivotilla tehty histogrammi naisten kuukausituloista manuaalisella luokittelulla.



Kuva B.4. Pivotilla tehty pisteparvi miesten kuukausituloista ja ympäristöhuolestuneisuudesta.



Kuva B.5. Pivotilla tehty pisteparvi naisten kuukausituloista ja ympäristöhuolestuneisuudesta.